Big Data for Measuring the Information Society: Country Report – Colombia

Contents

1. BACKGROUND AND CONTEXT

1.1. PROJECT DESCRIPTION

The aim of the project is to use big data from the telecommunication industry to improve and complement existing statistics and methodologies to measure the information society. The results of this project are expected to help countries and ITU to produce official statistics on information and communication technology (ICT) and to develop new methodologies by combining new and existing data sources. This will directly benefit policy-makers who will have access to new official statistics and benchmarks. It will also benefit data producers at the national level by guiding them in the use of big data for ICT measurement.

Specifically, the project:

- Combines the role of ITU as a standard-setting organisation in terms of global ICT measurement with official data producers' experience and interest in working with big data. Those data producers who can and are willing to share big data were invited to join the project to explore new methodologies and data sources to produce official ICT statistics. ICT data producers include: national statistical offices, telecommunication regulatory authorities, Mobile Network Operators (MNO), and Internet Service Providers (ISP).

- Takes advantage of ITU's involvement and experience in existing big data initiatives: ITU has organized a number of big data sessions in terms of regulation and monitoring, published information on the opportunities and challenges of big data, and developed a big data project on mobile phone data for health. ITU is also an active member of the UN Global Working Group on Big Data for Official Statistics, has networked with public and private organizations working on the topic, and has identified a number of potential (public and private) partners that were interested and benefitted from participating in this project.

- Explores and analyses the kind of new information society data and statistics that ITU and other stakeholders (including policy-makers, analysts and other data producers and users) would value most, and which are currently not existing.

- Engages with ICT data producers to discuss the development of specific indicators for new statistics that could be produced through big data analytics, and by combing new and existing data sources and methodologies.

The first phase of the project involves several country pilot studies, which serve as concrete examples of how big data from the ICT industry (MNOs, ISPs, etc.) can be used and/or combined with existing official data (from ITU, from national statistical offices and regulators) to produce new statistics, benchmarks and methodologies to measure the information society. These pilot studies cover different types of big data sources (e.g. MNOs, ISPs), and several ITU regions.

1.2. CONTEXT PILOT COUNTRY

REPUBLIC OF COLOMBIA

Table 1: Main social, economic and ICT indicators

SOCIAL AND ECONOMIC CONTEXT	2010	2015
Population (in millions)	46	48
Urban population (% of total)	75	76
GNI per capita (USD)	5 540	7 130
GDP growth, 2005–2010 and 2010–2015 (annual average in %)	4.4	4.6
Adult literacy rate (% ages 15 and older)	93	95
Gross primary, secondary and tertiary school enrolment (%)	89	88
Fixed-telephone subscriptions (per 100 inhabitants)	15.5	14.4
Mobile-cellular telephone subscriptions (per 100 inhabitants)	95.8	115.7
Fixed-broadband subscriptions (per 100 inhabitants)	5.7	11.2
Households with a computer (%)	26.1	45.5
Households with an Internet connection (%)	19.3	41.8

Source: The Little Data Book on Information and Communication Technology 2017. Available at http://www.itu.int/en/ITU-D/Statistics/Documents/publications/ldb/LDB_ICT_2017.pdf.

The ICT sector represents a considerable part of the country's GDP, reaching 7.5 per cent in 2015 (Table 1).[1] The mobile market is one of the most prosperous within the region, and still has immense potential. The Government is actively engaged in educating the population to make use of telecommunications services, investing in programmes to provide digital education.

Mobile services: The mobile market is dominated by three main operators: Comcel, Movistar and Tigo. There is increasing competition by new entrants, mobile virtual network operators (MVNOs), such as Uff! Móvil, ETB, Móvil Exito, Avantel and Virgin Mobile. From the entry of the MVNOs in the market in 2010, they had accumulated 6 per cent market share by 2015. Mobile use for Internet access has intensified over recent years. In 2016, 70 per cent of Internet users stated they used mobile devices to access the Internet.[2] There is potential for further mobile services expansion as the mobile-broadband penetration remains relatively low compared with the regional average. Additionally, mobile services are a welcome solution to the underdeveloped fixed-line infrastructure, especially in the remote areas.[3]

[1] Telecoms & IT, see: https://www.oxfordbusinessgroup.com/colombia-2016/telecoms-it

[2] National Statistical Institute – DANE, see: http://www.dane.gov.co/index.php/estadisticas-por-tema/tecnologia-e-innovacion/tecnologias-de-la-informacion-y-las-comunicaciones-tic/indicadores-basicos-de-tic-en-hogares

[3] Colombia - Mobile Infrastructure, Broadband, Operators - Statistics and Analyses, see: https://www.budde.com.au/Research/Colombia-Mobile-Infrastructure-Broadband-Operators-Statistics-and-Analyses?r=51.

Competition and service quality in this segment are expected to improve following the auction of frequencies in the 700 MHz and 1 900 MHz bands later in 2017.[4]

Fixed services: Colombia's fixed-telephony has been slower to lose ground to the mobile services than to regional and global trends. This can be partly attributed to the fact that fixed-telephony services are not subject to the value-added tax whereas mobile services are, which results in more affordable services. Fixed-broadband is mainly supplied via cable technology, with DSL representing 40 per cent of market share. Fibre-based technology remained at 5 per cent as of 2016, owing to its dependence on the development of fibre-infrastructure.[5]

Government policy: The government body in charge of overseeing the telecommunications sector, the *Comisión de Regulación de Comunicaciones*, has played an increasingly important part in the sector development. Its current action plan, the *Agenda Regulatoria 2017–2018*, promotes sector development through fostering competition and innovation, improving service quality and reducing prices, which at present are below the regional and global averages.[6] Another important programme, the *Plan Vive Digital para la Gente 2014–2018*, led by the *Ministerio de Tecnologías de la Informacíon y las Comunicaciones*, concentrates on leveraging the ICT development to positively impact the society, in terms of job generation, poverty reduction and digital education, among others.[7]

Conclusion: Colombia's telecommunications sector has been flourishing over recent years following public and private efforts. There remains a significant need to develop infrastructure, in particular regarding fixed-broadband. The numerous governmental programmes have encountered significant success with the increase of service penetration, especially within the lower socio-economic groups.

In terms of the country's political-territorial organization, the 1991 Constitution established 32 departments, divided into 1 101 municipalities and 20 departmental *corregimientos* (smaller than municipalities), which are listed in Colombia's *División Política y Administrativa* (Divipola).

DIVISIÓN POLÍTICA Y ADMINISTRATIVA OF COLOMBIA (DIVIPOLA)

Divipola is an encoding system used to maintain an organized and up-to-date list of all the units into which the territory of Colombia is divided, ensuring maximum stability for the identification of each department, municipality, departmental *corregimiento* and town (Figure 1).

Divipola is periodically updated by the National Administrative Department of Statistics (*Departamento Administrativo Nacional de Estadística*, or DANE) to reflect changes in territorial organization, using information submitted by the municipal and departmental administrations. It is a source of reference on the country's administrative and political organization.[8]

[4] MinTIC seeks expressions of interest in 700MHz, 1900MHz auction, see: https://www.telegeography.com/products/commsupdate/articles/2017/03/28/mintic-seeks-expressions-of-interest-in-700mhz-1900mhz-auction/.
[5] Official Website of Statistics of the ICT Sector - Colombia TIC, see: http://colombiatic.mintic.gov.co/602/w3-channel.html
[6] See: CRC (2016).
[7] El Plan Vive Digital 2014-2018, http://www.mintic.gov.co/portal/vivedigital/612/w3-article-19654.html.
[8] See *División Política y Administrativa de Colombia*, available at: http://geoportal.dane.gov.co:8084/Divipola/.

Figure 1 : División política y administrativa of Colombia (divipola)

Source: Departamento Administrativo Nacional de Estadistica -DANE (http://geoportal.dane.gov.co:8084/Divipola/).

1.3. PILOT PROJECT TIMELINE

Table 2: Project timeline

Activity	Date
Project inauguration.	08/09/2016
Meeting to present the project to network and fixed and mobile service providers.	09/09/2016
Working meeting with network and service providers to select the data-processing model and discuss the need to postpone data access and processing for 2017.	02/11/2016
Dispatch to each information provider of a request for a formal commitment to the project, selection of the data-processing model, estimate of data volume and projected date of access.	03/11/2016
Comunicación Celular S.A. (Comcel), formally confirms its participation in the project, selecting processing model 2 (data processed by the Ministry of Information and Communication Technologies (MinTIC) and DANE), with a projected date of access to the information of 15 April 2017 and a data volume of 5 terabytes. Providers TigoUne and Movistar do not confirm their participation in the project.	17/11/2016

Activity	Date
Meeting between MinTIC and DANE to define the architecture and identify the technological partner for configuring the data centre needed to develop the pilot project in Colombia.	20/12/2016
MinTIC develops the first draft confidentiality agreement linking the parties to the project and covering data access.	20/01/2017
MinTIC, DANE and Comunicación Celular S.A. (Comcel), start the legal review of the confidentiality agreement.	24/02/2017
The confidentiality agreement is signed by: (i) MinTIC; (ii) DANE; and (iii) Comunicación Celular S.A. (Comcel).	28/08/2017
MinTIC and DANE process the the data and produce the indicators	August-September 2017
MinTIC, DANE and ITU produce the project report	October-November 2017

2. STAKEHOLDERS IN THE PROJECT

Table 3: Main stakeholders

Entity	Nature	Role in the project
Ministry of Information and Communication Technologies (MinTIC)	Public	Coordinator (Colombia focal point) and facilitator.
National Administrative Department of Statistics (DANE), the national statistical office of Colombia.	Public	Facilitator, technological partner providing pilot project infrastructure (data centre).
Comunicación Celular S.A. (Comcel).	Private	Network and fixed and mobile service provider. Data provider

3. GETTING ACCESS TO THE DATA: PROCEDURES, LEGAL DOCUMENTS AND CHALLENGES

3.1. DATA PROCESSING TIMELINE

Table 4: Data processing timeline

Activity	Date
Dispatch of concept data (sample data) from Comcel S.A. to MinTIC and DANE for the analysis of the fields required for the project.	04/04/2017
Configuration of the virtual private network for SFTP data transfers between Comcel and DANE.	22/05/2017
Trial data made available for the first quarter of 2017.	24/05/2017
Final trial data made available for the first quarter of 2017.	15/06/2017
The provider (Comcel) starts transferring official second-quarter data to DANE.	31/08/2017

Activity	Date
The provider (Comcel) finishes transferring official second-quarter data to DANE.	15/09/2017
Start of data processing.	18/09/2017
Start of the calculation of pilot project indicators.	25/09/2017
An inconsistency is spotted in the information provided, owing to the fact that one antenna has a code exceeding the anticipated 6 characters for the field; the consequent loss of records has a direct impact on the calculation of indicators.	05/10/2017
Follow-up meeting with the provider, Comcel, to validate and adapt the data.	12/10/2017
Start of retransmission of 2017 second-quarter official data from the provider (Comcel) to DANE, with the provider Comcel S.A. sending summary tables for the calculation of indicators	18/10/2017
Restart of the calculation of pilot project indicators	19/10/2017

3.2. LEGAL DOCUMENTS AND CHALLENGES

In order for the Colombia pilot project to obtain the anticipated results, Comunicación Celular S.A. (Comcel), which was the sole information provider involved in the project and had a 49% share in the mobile voice market in the fourth quarter of 2016,[9] had to provide confidential technical information on its business strategies, such as the location of its antennas and the type of technology they used; it also had to provide information on network events contained in the call detail record (CDR).

As a condition of its participation in the pilot project, Comunicación Celular S.A. (Comcel) therefore demanded that the entities that would have direct access to the raw data – in this case, DANE and MinTIC – sign an agreement of confidentiality and non-disclosure of classified or reserved information.[10]

In addition, the officials from the public entities participating in the pilot project (DANE and MinTIC) who would have access to Comcel S.A. data agreed to enter into an "undertaking for access to classified information", given the sensitive nature of the information transmitted by Comcel S.A. to DANE, a public entity participating in the pilot project as a provider of technological infrastructure for storing and calculating indicators.

The model undertaking signed by DANE and MinTIC officials in order to obtain access to the data is set out below.

UNDERTAKING FOR ACCESS TO CONFIDENTIAL INFORMATION
Big Data for Measuring the Information Society pilot project of the International Telecommunication Union (ITU)
Bogotá, date XXXXX

[9] Official portal for ICT sector statistics, available at http://colombiatic.mintic.gov.co.

[10] Under one of the clauses of the confidentiality agreement signed between the parties involved in the Big Data for Measuring the Information Society pilot project, it was agreed that the model agreement would remain reserved and strictly confidential.

XXXXXXXXXXX, the holder of Citizenship Card No. XXXXXXX of XXXXX, in his/her capacity as an official/contractor of XXXXXXXXXXXXXXXXXXXXXX, declares that he/she is aware of the confidential and reserved nature of the information to which he/she will be granted access by DANE during the period between DAY MONTH 2017 and DAY MONTH 2017, whereby:

1. The information to which DANE will grant access shall be that delivered by Comcel in the framework of the ITU Big Data for Measuring the Information Society pilot project;

2. DANE shall grant access to the said information exclusively for the conduct of activities of which it has been informed by the official/contractor, which are described under his/her signature, and which are carried out in the exercise of his/her position or contractual obligations.

Consequently, in performance of the obligations set out in the law and the work contract, the MinTIC official/contractor undertakes to:

1. Fulfil the security conditions for the handling and storage of information in line with the provisions of existing institutional processes and regulations;

2. Respect the confidential nature of any information and documents that DANE provides and of which he/she has knowledge, in accordance with the applicable constitutional and legal rules, by virtue of which they shall in no case be published, distributed, reproduced, divulged, commented on or provided to other persons or entities by the official/contractor, be it in printed or electronic form, orally, or by any other means, with due regard in all cases for the terms of the confidentiality agreement reached between Comunicación Celular S.A. (Comcel S.A.), the Ministry of Information and Communication Technologies and the National Administrative Department of Statistics;

3. Use the information exclusively for the development of the project, and for no other purpose or research. Should the employee/contractor wish to use the information for another purpose, he/she shall advise accordingly and request prior authorization.

Lastly, the employee/contractor declares that he/she knows and agrees that failure to fulfil the obligations set out in this document will entail the civil, disciplinary, penal, fiscal and other consequences stipulated in the legislation in force.

_____

XXXXXXXXXX
C.C. No. XXXXX of XXXXXXX
Position: XXXXXXXXX
Department: XXXXXXXX
Telephone: +(57) XXXXXXXX

Purpose for which the information will be used:

Production and distribution of statistics and indicators proposed by ITU and DANE on information and communication technologies (ICT), based on the consultation, processing, design and analysis of mobile telephone call detail records (CDR).

_____

XXXXXXXXXXXX
Approved: XXXXXXXXXX

The main challenges encountered in the development of the big data pilot project were the following:

I. Selecting the single data-processing model
II. Bringing information providers into the project
III. Support staff for the construction of processing scripts
IV. Availability of data centres (main disadvantage: lack of storage space – up to 20 TB – owing to the use of the Hadoop Distributed File System, or HDFS)
V. Signing of a confidentiality agreement between the entities involved
VI. TAC database (references mobile equipment information using the IMEI number)
VII. Calculating the indicators – data of interest, owing to the processing time

3.3. RESOURCES REQUIRED FOR THE PROCESSING

The platform used to process the pilot project data consisted of four (4) virtual nodes (since the DANE platform was configured in this way, although it could be configured with independent physical servers), each one of which was configured as follows:

- Disc: 5 TB
- Memory: 24 GB
- CPU: Intel 8 cores x86
- Operating system: Oracle Linux

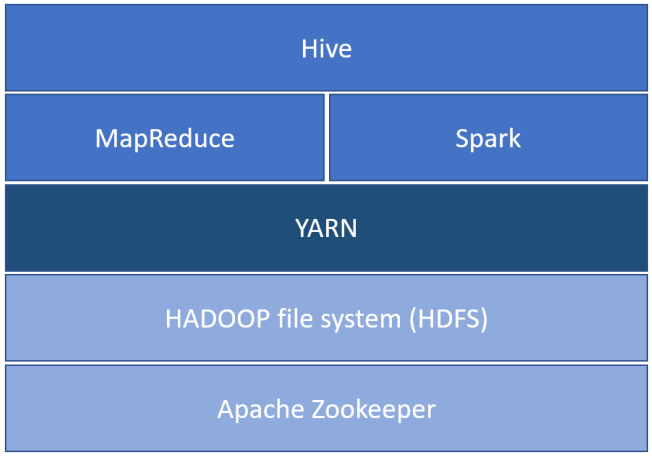
The four-node configuration was identified as the requisite minimum, given that this was a pilot project and that the resources available for the project were limited.

The disc size was based on the mobile-cellular service provider's announcement of a data volume of six (6) TB for the quarter to be processed, and took into account the Hadoop platform's default requirement of data triplication to ensure the benefits of multiprocessing and availability. This implied the need for at least 18 TB of storage; 20 TB were therefore made available in all, leaving an extra 2 TB for summary tables.

It is on this platform that the Hadoop Cloud CDH 5.11.0 cluster was installed.

Figure 2 presents the main components of the Hadoop platform that were used to develop the project. Tests were conducted with both Spark and MapReduce, but the results obtained using the latter were the same and in some cases a bit better than the former; this makes sense, given that Spark has significant advantages over MapReduce when it comes to processing information in real time, whereas this specific pilot project worked with batch information.

Figure 2: Components of the Hadoop platform



Source: Technical preparations – Big Data pilot project in Colombia.

The data intake process and platform configuration were extremely simple, as the process would be carried out only once during the pilot project and data from only one mobile-cellular service provider would be processed.

The mobile service provider made the files available remotely, using the **fuse-sshfs-2.5-1.el7.x86_64** package; the files were assembled as one package on the Hadoop platform's main server. Each file was subsequently transferred and entered directly into the HDFS, so as not to create additional storage needs.

4. METHODOLOGY, TECHNICAL DESCRIPTION, TOOLS

4.1. DESCRIPTION OF THE DATA USED IN THE PROJECT

Table 5: Descprition of data fields

PROVIDER	FIELD	DESCRIPTION	TYPE OF DATA	EXAMPLE DATUM AS DEFINED BY ITU
MNO NETWORK DATA	CELL_ID	Antenna ID	Text	1000003

PROVIDER	FIELD	DESCRIPTION	TYPE OF DATA	EXAMPLE DATUM AS DEFINED BY ITU
MNO NETWORK DATA	TECH_GEN	Technological generation	2G, 3G, LTE coded text (only one)	3G
MNO NETWORK DATA	LAT	WGS84 latitude of the antenna	Numerical	41,514441
MNO NETWORK DATA	LON	WGS84 longitude of the antenna	Numerical	46,689006
MNO NETWORK DATA	CELL_GEOM	Vector/polygon of the antenna coverage area		SHP file or other GIS format
MNO CDR/IPDR DATA	ID	Subscriber ID, based on the IMSI code	Text	90000002
MNO CDR/IPDR DATA	IMEI	IMEI code	Whole number	990007033671238
MNO CDR/IPDR DATA	TYPE_COMMERCIAL_PRIVATE	Type of contract	Coded numeral 1 – Residential 2 – Non-residential	1
MNO CDR/IPDR DATA	TYPE_PRE_PST	Type of payment	Coded numeral 1 – Prepaid 2 – Postpaid	1

PROVIDER	FIELD	DESCRIPTION	TYPE OF DATA	EXAMPLE DATUM AS DEFINED BY ITU
MNO CDR/IPDR DATA	TYPE_VOICE_DATA	Type of plan	Coded numeral 1 – Voice only 2 – Voice and data 3 – Data only	2

MNO CDR/IPDR DATA	TYPE_EVENT	Type of CDR/IP detail record (IPDR)	Coded numeral 1 – Outbound voice CDR, with no Internet use 2 – Inbound voice CDR, with no Internet use 3 – Outbound text message CDR (SMS, MMS), with no Internet use 4 – Inbound text message CDR (SMS, MMS), with no Internet use 5 – Other type of CDR with no Internet use and not calls or text messages (SSE, USSD, others) 6 – Open Internet IPDR, any Internet use except MMS messages 7 – Closed Internet IPDR, any Internet use except MMS messages	6

PROVIDER	FIELD	DESCRIPTION	TYPE OF DATA	EXAMPLE DATUM AS DEFINED BY ITU
MNO CDR/IPDR DATA	DATETIME	Start time and date of the CDR, IPDR (UTC time)	Timestamp YYYY-MM-DD hh:mm:ss	2016-05-13 10:15:23
MNO CDR/IPDR DATA	DURATION	Duration of the event in seconds	Numerical value	933
MNO CDR/IPDR DATA	TYPE_DOMESTIC_OUTBOUND_INBOUND	Type of record	Coded numeral 1 – National 2 – Inbound roaming 3 – Outbound roaming	1
MNO CDR/IPDR DATA	MCC	Mobile country code: SIM card country of origin, extracted from the IMSI code	Whole number	732
MNO CDR/IPDR DATA	CELL_ID	Antenna ID, only for national records and incoming roaming	Text	1000003
MNO CDR/IPDR DATA	DATA_VOLUME	Total volume data uploaded and downloaded during the connection, in bytes	Whole number	238840022

Source: Technical preparations – Big Data pilot project in Colombia.

Table 6: Example raw data from the MNO

TECH_GEN	LON	LAT	CELL_GEOM	ID	IMEI	TYPE_COMMERCIAL_PRIVATE	TYPE_PRE_PST	TYPE_VOICE_DATA	TYPE_EVENT	DATE_TIME	DURATION	TYPE_DOMESTIC_OUTBOUND_INBOUND	MCC
2G	-74.176.608	4.589.642		732101059416478	35671805876072	2	1	1	2	1/01/2017 14:11	729	1	732
3G	-741.533	45.908		732101017979207	35633405124004	1	2	2	2	1/01/2017 17:44	321	1	732
2G	-75.320.111	8.911.472		732101233305903	35467106533256	2	1	1	1	1/01/2017 0:06	44	1	732
3G	-738.259	561.222		732101089180444	1471000279787	1	2	2	2	1/01/2017 6:13	15	1	732
2G	-7.405.164	46.902		732101419745316	1285900916794	1	2	1	1	1/01/2017 12:33	54	1	732
3G	-7.586.835	519.846		732101180287408	35944106644017	1	1	1	1	1/01/2017 1:32	13	1	732
3G	-75.609.361	6.242.472		732101171477747	35443505514843	1	2	2	2	1/01/2017 18:58	33	1	732
3G	-744.135	42.005		732101407154170	35214107778259	1	1	1	2	1/01/2017 18:11	25	1	732
2G	-74.755.806	10.784.028		732101229174262	35255205648756	1	1	1	1	1/01/2017 21:38	310	1	732
LTE	-75.568.949	6.351.817		732101179383211	35383405602688	1	2	2	1	1/01/2017 8:13	313	1	732

Source: Sample data from the network and telecommunications service provider Comunicación Celular S.A. (Comcel).

4.2. REFERENCE DATA USED IN THE PROJECT

The reference data used in Colombia to develop the ITU Big Data for Measuring the Information Society pilot project are drawn chiefly from studies and/or research conducted by national entities that provided contrasting elements for the purpose of comparing the data and the results obtained after calculating the big data indicators.

Those studies and/or research can be classified by data-collection method (Table 7).

Table 7: Reference data used in the project

Censuses / surveys	Periodic information reports and/or administrative records
2005 National Census of the Population http://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion	Arrival of foreigners in Colombia http://migracioncolombia.gov.co/index.php/es/?option=com_content&view=article&id=718
Subject: Information Technology Survey of the Quality of Life http://www.dane.gov.co/index.php/estadisticas-por-tema/tecnologia-e-innovacion	Quarterly ICT Bulletin http://colombiatic.mintic.gov.co/602/w3-channel.html
División Político Administrativa de Colombia http://geoportal.dane.gov.co:8084/Divipola/	Mobile provider coverage reference map http://www.claro.com.co/personas/soporte/mapas-de-cobertura/

Source: Technical preparations – Big Data pilot project in Colombia.

Divipola has three levels of disaggregation, identifying Colombia's departments, municipalities and towns. A case study based on that breakdown is presented in Table 8 as an example.

Table 8: example of administrative divisions in Colombia

Department code	Municipality code	Town code	Department name	Municipality name	Town name
05	05001	05001000	ANTIOQUIA	MEDELLÍN	MEDELLÍN
05	05001	05001001	ANTIOQUIA	MEDELLÍN	PALMITAS
05	05001	05001004	ANTIOQUIA	MEDELLÍN	SANTA ELENA
05	05001	05001009	ANTIOQUIA	MEDELLÍN	ALTAVISTA

4.3. ACCESS TO THE DATA

For privacy protection reasons, and because the initial raw data from data providers can be considered to include confidential business information, there are three tiers or phases of data processing.

- Tier I – initial, raw, not aggregated data extracted by the data providers from their databases and registries, that is the basis for calculation and which may include private and confidential business information. This can also be referred to as microdata.
- Tier II – initially aggregated data per data provider with no private and some (or no) confidential business information, but which may still considered sensitive and not shared with third parties. This can also be referred to as macrodata.
- Tier III – Tier II aggregated indicators from different data providers merged that can be publicly shared and that do not include any private or confidential business information. Tier III represents the resulting statistical indicators of the project.

DATA INTAKE

An extremely simple mechanism was used for data intake (microdata), as the pilot project process would be carried out only once, using information from a single mobile-cellular service provider.

The source files were made available in compressed text format, with fields separated by the pipe character (|), one file for each day of information. Each file contained an average of one billion two hundred million records (1 200 000 000) and occupied about 35 GB of space.

The mobile service provider made the files available remotely, using the fuse-sshfs-2.5-1.el7.x86_64 package; the files were assembled as one package on the Hadoop platform's main server. Each file was subsequently transferred and entered directly into the HDFS, so as not to create additional storage needs, using the PUT instruction.

On average, and given that the transfers took place over a non-dedicated communication channel, each file took approximately two hours to transfer and incorporate into the HDFS.

4.4. PROCESSING THE DATA

QUALITY ASSURANCE

Because the project involved processing 104.5 billion records (104 545 252 173, to be exact), carrying out quality assurance procedures on the microdata might have required excessive time and resources that were not necessarily available. Quality assurance was therefore carried out in phases and never on all of the microdata.

Phase I – Quality check of a data sample

Initially, a data sample was used that had been sent by the mobile-cellular service provider and comprised approximately one hundred thousand (100 000) records. The sample was checked for the value domain of each field, both for categorical and continuous variables. Comments and adjustments were made accordingly.

Phase II – Quality check of data transfer and intake

The telephony service provider was asked to report the total number of records delivered in the daily files, and these were compared with the data received once the files had been transferred and entered into the HDFS platform and Hive. The size of each of the files transferred was also verified.

Phase III – Quality check using summaries of indicator raw information

Once the records had been reviewed for completeness, summaries started to be processed for the calculation of each indicator.

Initially, and given that the provider did not send a separate file on stations and antennas, the different combinations of latitude, longitude and technology had to be identified in order to identify panels and later stations. At the outset, 7 836 stations were identified, and that number was communicated to the provider for verification.

Subsequently, intermediate tables were constructed in Hive to calculate each indicator, and control totals were calculated of, for example:

- the total number of events
- the total duration of events
- the total records.

CALCULATION OF INDICATORS ON THE HADOOP PLATFORM

Hive processing was deemed fundamental for the calculation of indicators on the Hadoop platform, as it offers major advantages that facilitate processing of the types of indicator proposed by ITU. It is nonetheless important to bear in mind certain fundamental elements in terms of tuning the platform, to ensure reasonable processing times given the minimal platform available for the project.

The following steps were taken once the 91 files for the quarter had been incorporated into the Hadoop platform in the original compressed text format made available by the mobile-cellular service provider.

1. A directory was created in the HDFS structure to host the database.
2. The Hive database was created.
3. An external CDR table was created (Table 9), divided by year, month and day, using the text files made available by the mobile-telephony service provider.

Table 9: Description of the CDR table created in Hive

FIELD	TYPE	DESCRIPTION
tec	string	TECH_GEN
lon	double	LONGITUDE
lat	double	LATITUDE
geo	string	CELL_GEOM
imsi	bigint	ID (IMSI)
imei	bigint	IMEI

FIELD	TYPE	DESCRIPTION
com_pri	int	TYPE_COMMERCIAL_PRIVATE
pre_pst	int	TYPE_PRE_PST
voi_dat	int	TYPE_VOICE_DATA
typ_eve	int	TYPE_EVENT
dat_tim	string	DATE TIME
dur	bigint	DURATION
typ_dom	int	TYPE_DOMESTIC_OUTBOUND_INBOUND
mcc	int	MCC
año	int	
mes	int	
dia	int	
# Partition Inf		
# col_name	data_type	comment
año	int	
mes	int	
dia	int	

Source: Technical preparations – Big Data pilot project in Colombia.

4. The CDR table was loaded (this did not imply duplication of storage resources) with the text files made available by the mobile-cellular service provider.

5. The number of daily records on the CDR table was calculated and compared for quality control purposes.

6. ANTENNA_PANEL_PERIOD, ANTENNA_PANEL and ANTENNADEF tables were created in orc.compress format, using the distinct combinations of latitude, longitude and technology; antennas were localized geographically using latitude and longitude (GIS) and the final table adjusted (Table 10).

Table 10: Description of the table with information on the antennas created in Hive

FIELD	TYPE	DESCRIPTION
id	int	Antenna identity (generated)
lat	double	Latitude
lon	double	Longitude
tec	string	Interlinkage of technologies available on the antenna
tec_id	int	Identification of technologies available on the antenna
tec_num	int	Number of technologies available on the antenna
divipola	string	Complete dept-municipality-town code
dep	string	Department
mun	string	Municipality
cpo	string	Town
clase	string	Urban/rural type

Source: Technical preparations – Big Data pilot project in Colombia.

7. The CDR_RES table was created in orc.compress format, divided by year and month, as the basis for indicators comprising one month of information or more (Table 11).

Table 11: Description of the CDR_RES table created in Hive

FIELD	TYPE	DESCRIPTION
tec	string	TECH_GEN
lon	double	LONGITUDE
lat	double	LATITUDE
imsi	bigint	ID (IMSI)
imei	bigint	IMEI
com_pri	int	TYPE_COMMERCIAL_PRIVATE

FIELD	TYPE	DESCRIPTION
pre_pst	int	TYPE_PRE_PST
voi_dat	int	TYPE_VOICE_DATA
typ_eve	int	TYPE_EVENT
typ_dom	int	TYPE_DOMESTIC_OUTBOUND_INBOUND
mcc	int	MCC
dur	bigint	DURATION
tot	bigint	TOTAL EVENTS
ano	int	
mes	int	
# Partition Info		
# col_name	data_type	comment
ano	int	
mes	int	

Source: Technical preparations – Big Data pilot project in Colombia.

8. Other summary tables were created specifically to calculate specific inidactors.

Table 12: Description of the table *with the information on indicator BD03* created in Hive

FIELD	TYPE	DESCRIPTION
trim	int	QUARTER
ope	string	OPERATOR
divipola	string	DIVIPOLA
dep	string	DEPARTMENT
mun	string	MUNICIPALITY

FIELD	TYPE	DESCRIPTION
clase	string	CLASS
tec	string	TECH_GEN
des_eve	string	DESCRIPTION EVENTS
dom	string	DOMESTIC
dur	bigint	TOTAL DURATION
tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

Table 13: Description of the table with the information on indicator BD04 created in Hive

FIELD	TYPE	DESCRIPTION
trim	int	QUARTER
ope	string	OPERATOR
divipola	string	DIVIPOLA
dep	string	DEPARTMENT
mun	string	MUNICIPALITY
clase	string	CLASS
tec	string	TECH_GEN
des_eve	string	DESCRIPTION EVENTS
dom	string	DOMESTIC
dur	bigint	TOTAL DURATION
tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

INDICATOR BD05 – Four tables at national, departmental, municipal and class level.

Table 14: Description of the table with the information on indicator BD05 created in Hive, class level

FIELD	TYPE	DESCRIPTION
divipola	string	DIVIPOLA
dep	string	DEPARTMENT
mun	string	MUNICIPALITY
clase	string	CLASS
tec	string	MAXIMUM TECHNOLOGY
num	int	NUMBER OF DIFFERENT IMSI

Source: Technical preparations – Big Data pilot project in Colombia.

Table 15: Description of the table with the information on indicator BD05 created in Hive, national level

FIELD	TYPE	DESCRIPTION
tec	string	TECHNOLOGY
num	int	NUMBER OF DIFFERENT IMSI

Source: Technical preparations – Big Data pilot project in Colombia.

INDICATOR BD06 – Four tables at national, departmental, municipal and class level.

Table 16: Description of the table with the information on indicator BD06 created in Hive, class level

FIELD	TYPE	DESCRIPTION
divipola	string	DIVIPOLA
dep	string	DEPARTMENT
mun	string	MUNICIPALITY
clase	string	CLASS

com_pri	int	TYPE_COMMERCIAL_PRIVATE
pre_pst	int	TYPE_PRE_PST
voi_dat	int	TYPE_VOICE_DATA
num	int	NUMBER OF DIFFERENT IMSI
dur	bigint	TOTAL DURATION
tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

Table 17: Description of the table with the information on indicator BD06 created in Hive, departmental level

FIELD	TYPE	DESCRIPTION
dep	string	DEPARTMENT
com_pri	int	TYPE_COMMERCIAL_PRIVATE
pre_pst	int	TYPE_PRE_PST
voi_dat	int	TYPE_VOICE_DATA
num	int	NUMBER OF DIFFERENT IMSI
dur	bigint	TOTAL DURATION
tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

INDICATOR BD08 – Four tables at national, departmental, municipal and class level.

Table 18: Description of the table with the information on indicator BD08 created in Hive, class level

FIELD	TYPE	DESCRIPTION
trim	int	QUARTER
ope	string	OPERATOR

divipola	string	DIVIPOLA
dep	string	DEPARTMENT
mun	string	MUNICIPALITY
clase	string	CLASS
tac	string	8-DIGIT TAC IMEI
num	int	TOTAL EQUIPMENT
dur	bigint	TOTAL DURATION
tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

Table 19: Description of the table with the information on indicator BD08 created in Hive, departmental level

FIELD	TYPE	DESCRIPTION
trim	int	QUARTER
ope	string	OPERATOR
dep	string	DEPARTMENT
tac	string	8-DIGIT TAC IMEI
num	int	TOTAL EQUIPMENT
dur	bigint	TOTAL DURATION
Tot	bigint	TOTAL EVENTS

Source: Technical preparations – Big Data pilot project in Colombia.

5. RESULTS DERIVED FOR THE BIG DATA INDICATORS

The development in Colombia of the ITU Big Data for Measuring the Information Society pilot project served to calculate the big data indicators that are based on the records of network and mobile telecommunication service providers, specifically CDR data; the priority 1 mobile indicator BD12 (mobile international broadband traffic, by contract type) was not included in the calculations, because the provider, Comunicación Celular S.A. (Comcel), did not use the fields

required in the methodological handbook developed by ITU for calculating the indicator to manage and capture information on its network.

The pilot project analysed the mobile provider's CDR data for the second quarter of 2017 (April, May and June); it also used demographic information provided by DANE for the calculation specifically of priority 1 indicator BD02 (percentage of the population covered by a mobile-cellular network, by technology).

A summary list of the pilot project indicators is set out in Table 20, together with each indicator's calculation status.

Table 20: List of indicators calculated in the pilot project.

Indicator	Priority	Description	Status
BD01	2	Percentage of the land area covered by a mobile-cellular network, by technology	Not calculated
BD02	1	Percentage of the population covered by a mobile-cellular network, by technology	Not calculated
BD03	2	Usage of mobile-cellular networks for non-IP-related activities, by technology	Calculated Documented
BD04	1	Usage of mobile-cellular networks for Internet access, by technology	Calculated Documented
BD05	1	Number of subscriptions with access to technology	Calculated Documented
BD06	2	Active mobile voice and broadband subscriptions, by contract type	Calculated Documented
BD07	2	Average number of active mobile subscriptions per day, by contract type	Calculated Documented
BD08	1	Active mobile devices	Not calculated
BD09	2	IMEI conversion rate	Not calculated
BD10	1	Domestic fixed-broadband traffic, by speed, contract type	Not calculated
BD11	1	Domestic mobile-broadband traffic, by contract type, technology	Calculated Documented
BD12	1	International mobile-broadband traffic, by contract type	Not calculated
BD13	2	Inbound roaming subscriptions per foreign tourist	Calculated Documented
BD14	2	Fixed-broadband subscriptions, by technology	Not calculated
BD15	1	Fixed-broadband subscriptions, by speed	Not calculated
BD16	2	Classification of municipalities based on their daily activity	Theoretical description
BD17	2	Classification of areas as residential or non-residential	Theoretical description
BD18	2	Calculation of activity densities disaggregated by area	Theoretical description
BD19	2	Deployment of mobile network infrastructure by population density	Calculated Documented

BD01: Percentage of the land area covered by a mobile-cellular network, by technology

Not calculated

BD02: Percentage of the population covered by a mobile-cellular network, by technology

Not calculated

BD03: Usage of mobile-cellular networks for non-IP-related activities, by technology

This indicator was calculated using a summary table[11] structured as follows:

DATE, DDMMM,[12] URBAN/RURAL, TECH_GEN, TYPE_EVENT, TYPE_DOMESTIC_OUTBOUND_INBOUND, TYPE_COMERCIAL_PRIVATE, TOTAL EVENTS, TOTAL DURATION/DATA CONSUMPTION, UNIQUE_USERS.

The input table had 3 253 424 records for the ten variables identified.

To calculate the indicator, the events count was bundled by type of technology and geographical level. In each case, the indicator is the use of a type of technology (2G, 3G) for non-Internet-related activities expressed as a percentage of total network use.

The indicator was calculated in Phyton and R.

The entire country, urban and rural, 2G: 42.8%

The entire country, urban and rural, 3G: 57.2%

The entire country, rural, 2G: 60.1%

The entire country, rural, 3G: 39.9%

The entire country, urban, 2G: 37.2%

The entire country, urban, 3G: 62.8%

BD04: Usage of mobile-cellular networks for Internet access, by technology

This indicator was calculated using a summary table structured as follows:

DATE, DDMMM, URBAN/RURAL, TECH_GEN, TYPE_EVENT, TYPE_DOMESTIC_OUTBOUND_INBOUND, TYPE_COMERCIAL_PRIVATE, TOTAL EVENTS, TOTAL DURATION/DATA CONSUMPTION, UNIQUE_USERS.

The input table had 3 253 424 records for the ten variables identified.

To calculate the indicator, the events count was bundled by type of technology and geographical level. In each case, the indicator is the use of a type of technology (2G, 3G, LTE) for Internet-related activities expressed as a percentage of total network use.

The indicator was calculated in Phyton and R.

The entire country, urban and rural, 2G: 4.9%

[11] Indicators BD03, BD04 and BD011 were calculated using a single summary table to simplify data processing.

[12] DD refers to the department code, and MMM to the municipality code.

The entire country, urban and rural, 3G: 43.7%

The entire country, urban and rural, LTE: 51.4%

The entire country, rural, 2G: 13.5%

The entire country, rural, 3G: 70.9%

The entire country, rural, LTE: 15.6%

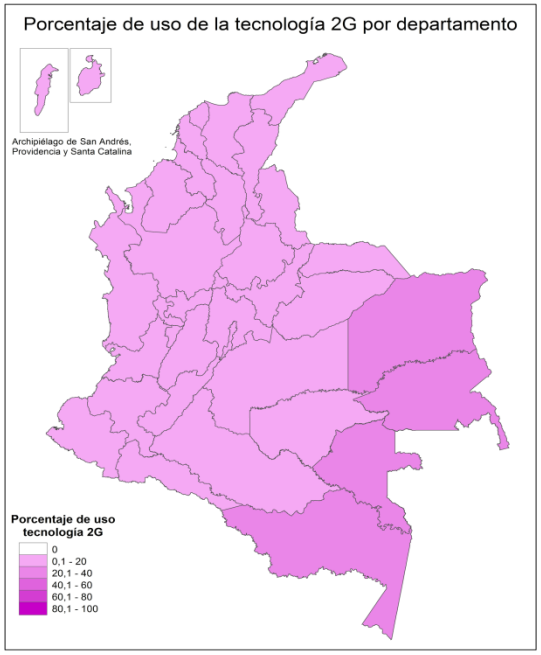
The entire country, urban, 2G: 3.3%

The entire country, urban, 3G: 38.8%

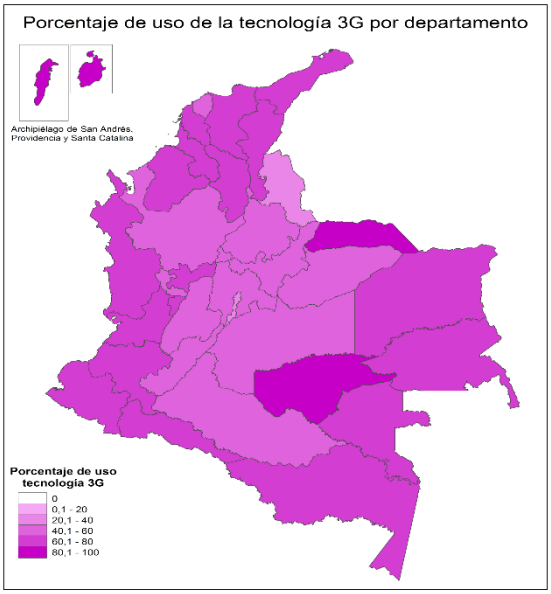
The entire country, urban, LTE: 57.9%

The results by department are set out in Map 1, Map **2** and Map ***3*** respectively for 2G, 3G and LTE networks.

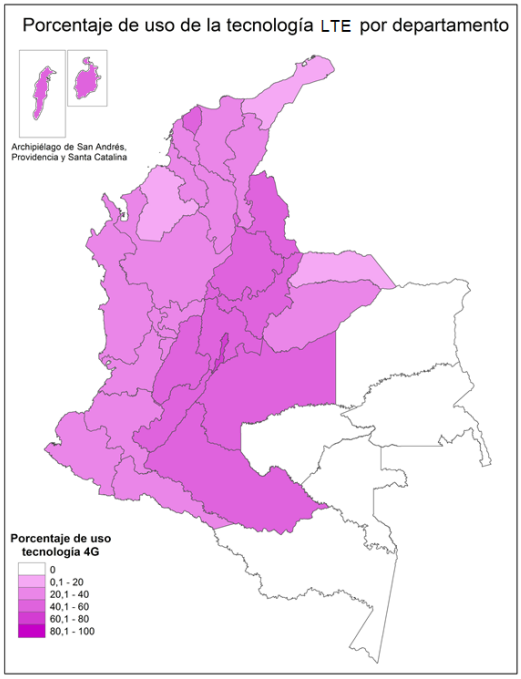
Map 1: Use of the 2G network for Internet access



Map 2: Use of the 3G network for Internet access



Map 3: Use of the LTE network for Internet access



BD05: NUMBER OF SUBSCRIPTIONS WITH ACCESS TO TECHNOLOGY

This indicator was calculated using a summary table structured as follows:

DDMMM, URBAN/RURAL, TECH_GEN, IMSI, TYPE_COMMERCIAL_PRIVATE, TYPE_PRE_PST, TYPE_VOICE_DATA.

To calculate the indicator, unique users were identified by technology on the mobile network and broken down by geographical level.[13] The national results are set out in Table 21 and the results by department in Table 22.

Table 21: BD05 results at the national level

Department	2G subscriptions	3G subscriptions	LTE subscriptions
National total	9 254 641	16 532 633	4 536 962

[13] For the purposes of the pilot project, the outcome value was listed after the indicator had been calculated. The indicator's calculation will nonetheless be analysed in greater depth, as the outcome figure is not the same as that obtained from administrative records.

Table 22: BD05 results at the departmental level

Department	2G subscriptions	3G subscriptions	LTE subscriptions
ANTIOQUIA	1 301 935	2 934 543	775 207
ATLÁNTICO	367 995	532 322	180 418
BOGOTÁ, D. C.	1 653 551	4 234 192	1 962 169
BOLÍVAR	470 631	497 710	69 310
BOYACÁ	395 673	574 216	105 111
CALDAS	283 067	558 734	121 803
CAQUETÁ	111 911	194 139	50 578
CAUCA	443 143	715 404	137 720
CESAR	261 267	306 486	45 722
CÓRDOBA	398 261	468 417	46 737
CUNDINAMARCA	404 638	571 580	90 084
CHOCÓ	90 401	148 254	25 036
HUILA	258 063	379 642	68 239
LA GUAJIRA	112 658	92 907	10 417
MAGDALENA	147 099	138 940	19 197
META	213 613	357 710	96 091
NARIÑO	279 886	383 114	63 878
NORTE DE SANTANDER	311 356	486 766	136 966
QUINDÍO	142 009	325 254	68 469
RISARALDA	154 075	333 718	50 771
SANTANDER	302 598	430 146	100 138
SUCRE	152 277	118 435	10 129
TOLIMA	249 562	345 989	51 165
VALLE DEL CAUCA	501 298	1 029 591	209 035
ARAUCA	56 209	87 397	3 513
CASANARE	77 581	125 532	28 534
PUTUMAYO	61 991	94 661	7 946
ARCHIPELAGO OF SAN ANDRÉS, PROVIDENCIA AND SANTA CATALINA	3 136	9 814	2 579
AMAZONAS	18 833	21 556	0
GUAINÍA	7 905	7 193	0
GUAVIARE	8 809	14 425	0
VAUPÉS	3 813	4 213	0
VICHADA	9 397	9 633	0

BD06: ACTIVE MOBILE VOICE AND BROADBAND SUBSCRIPTIONS, BY CONTRACT TYPE

This indicator was calculated using a summary table structured as follows:

DDMMM, URBAN/RURAL, TECH_GEN, IMSI, TYPE_COMMERCIAL_PRIVATE, TYPE_PRE_PST, TYPE_VOICE_DATA.

To calculate the indicator, unique subscriptions were identified by type of contract on the network.[14] The results at the national level are set out in Table 23.

Table 23: BD06 results at the national level

Class	Type of contract	Type of payment	Type of plan	Subscriptions
Urban	Residential	Prepaid	All[15]	27 429 033
Urban	Residential	Postpaid	Voice	184 023
Urban	Residential	Postpaid	Voice and data	728 851
Urban	Residential	Postpaid	Data	42 657
Urban	Non-residential	Prepaid	Voice	1 449 533
Urban	Non-residential	Postpaid	Voice	34 381
Urban	Non-residential	Postpaid	Voice and data	24 431
Urban	Non-residential	Postpaid	Data	340 745
Rural	Residential	Prepaid	All	1 664 880
Rural	Residential	Postpaid	Voice	12 111
Rural	Residential	Postpaid	Voice and data	45 591
Rural	Residential	Postpaid	Data	5 546
Rural	Non-residential	Prepaid	All	185 730
Rural	Non-residential	Postpaid	Voice	2 962
Rural	Non-residential	Postpaid	Voice and data	1 528
Rural	Non-residential	Postpaid	Data	19 978

BD07: AVERAGE NUMBER OF ACTIVE MOBILE SUBSCRIPTIONS PER DAY, BY CONTRACT TYPE

This indicator was calculated using a summary table structured as follows:

DDMMM, URBAN/RURAL, DATE, TYPE_COMMERCIAL_PRIVATE, TYPE_PRE_PST, TYPE_VOICE_DATA, NUMBER UNIQUE IMSI.

To calculate the indicator, unique subscriptions were identified by type of contract, payment mode and type of plan.[16] The preliminary results at the national level are set out in Table 24. Given the low value of subscriptions used per day (considering a total of more than 27 million active subscriptions in the reference period for Comcel), further validation would be necessary to produce this indicator.

[14] Given that Comunicación Celular S.A. (Comcel) is the only provider participating in the pilot project, the outcome information is aggregated at national level.

[15] Prepaid subscriptions were not possible to break down by type of plan (voice/data/voice and data).

[16] See note 14 above.

Table 24: BD07 results at the national level, by type of contract

Type of contract	Average activations per day
Non-residential	30 540
Residential	62 696
Postpaid	27 031
Prepaid	105 379
Data	3 581
Voice	61 217
Voice and data	60 458

BD08: ACTIVE MOBILE DEVICES

Not calculated

BD09: IMEI CONVERSION RATE

Not calculated

BD10: DOMESTIC FIXED-BROADBAND TRAFFIC, BY SPEED, CONTRACT TYPE

Not calculated

BD11: DOMESTIC MOBILE-BROADBAND TRAFFIC, BY CONTRACT TYPE, TECHNOLOGY

This indicator was calculated using a summary table structured as follows:

DATE, DDMMM, URBAN/RURAL, TECH_GEN, TYPE_EVENT, TYPE_DOMESTIC_OUTBOUND_INBOUND, TYPE_COMERCIAL_PRIVATE, TOTAL EVENTS, SUM DURATION/DATA CONSUMPTION, UNIQUE_USERS.

The input table had 3 253 424 records for the ten variables identified.

The indicator was calculated in Phyton and R.

The entire country, urban and rural, all generations: 50.3[17] (exabytes)

The entire country, rural, all generations: 9.4 (exabytes)

The entire country, urban, all generations: 40.9 (exabytes)

The entire country, urban and rural, residential, all generations: 43.0 (exabytes)

The entire country, urban and rural, non-residential, all generations: 7.3 (exabytes)

The entire country, urban and rural, 2G: 0.7 (exabytes)

[17] For the purposes of the pilot project, the outcome value was listed after the indicator had been calculated. The indicator's calculation will nonetheless be analysed in greater depth, as the outcome figure is not the same as that obtained from administrative records.

The entire country, urban and rural, 3G: 24.8 (exabytes)

The entire country, urban and rural, LTE: 24.8 (exabytes)

BD12: INTERNATIONAL MOBILE-BROADBAND TRAFFIC, BY CONTRACT TYPE

Not calculated

BD13: INBOUND ROAMING SUBSCRIPTIONS PER FOREIGN TOURIST

For this indicator, a summary table was generated with the variables DATE_MONTH, IMSI_UNIQUE, TYPE_INBOUND_ROAMING.

The count was subsequently established of unique subscriptions with types of roaming event, and the quotient was calculated for that value and the number of tourists entering the country during the months being studied.

The calculation was so simple that it was performed in Excel (Table 25).

Table 25: BD13 results

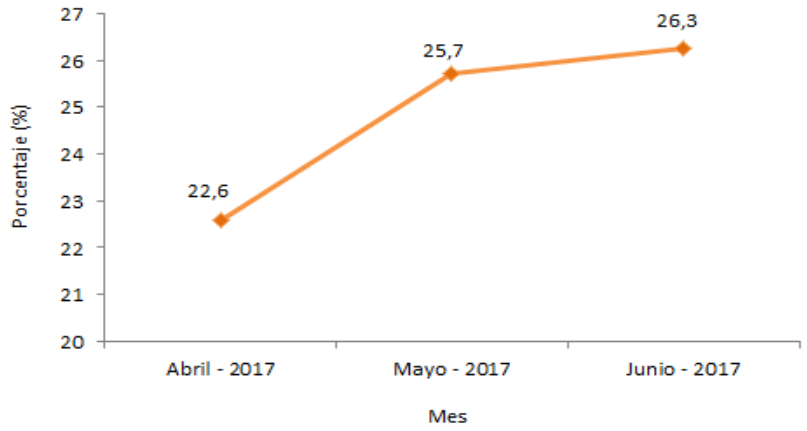
Month	Users of inbound roaming	Total incoming travellers*	Indicator (%)[18]
April 2017	114 600	507 346	22.6
May 2017	118 424	460 258	25.7
June 2017	124 408	473 571	26.3

Source: DANE - MinTIC calculations using COMCEL S.A. figures

* Includes: non-resident foreigners, Colombians residing abroad, number of passengers on international and cross-border cruises

[18] The sole source for inbound roaming data was Comunicación Celular S.A. (Comcel); the percentage calculated may therefore be skewed, as it does not include the remaining eight mobile providers offering services in Colombia.

Figure 3: Roaming use by foreign tourists



Source: Technical preparations – Big Data pilot project in Colombia.

BD14: FIXED-BROADBAND SUBSCRIPTIONS, BY TECHNOLOGY

Not calculated

BD15: FIXED-BROADBAND SUBSCRIPTIONS, BY SPEED

Not calculated

BD16: CLASSIFICATION OF MUNICIPALITIES BASED ON THEIR DAILY ACTIVITY

In the context of the Measuring the Information Society project, the statistics office (DANE) proposed to calculate four indicators based on daily call and data activity, which, supplemented with traditional statistical information, geographical information systems and administrative records, provide additional information for census purposes.

The input for those indicators is the summary table generated by the mobile operator of total events and their duration, by base station and at one-hour intervals.

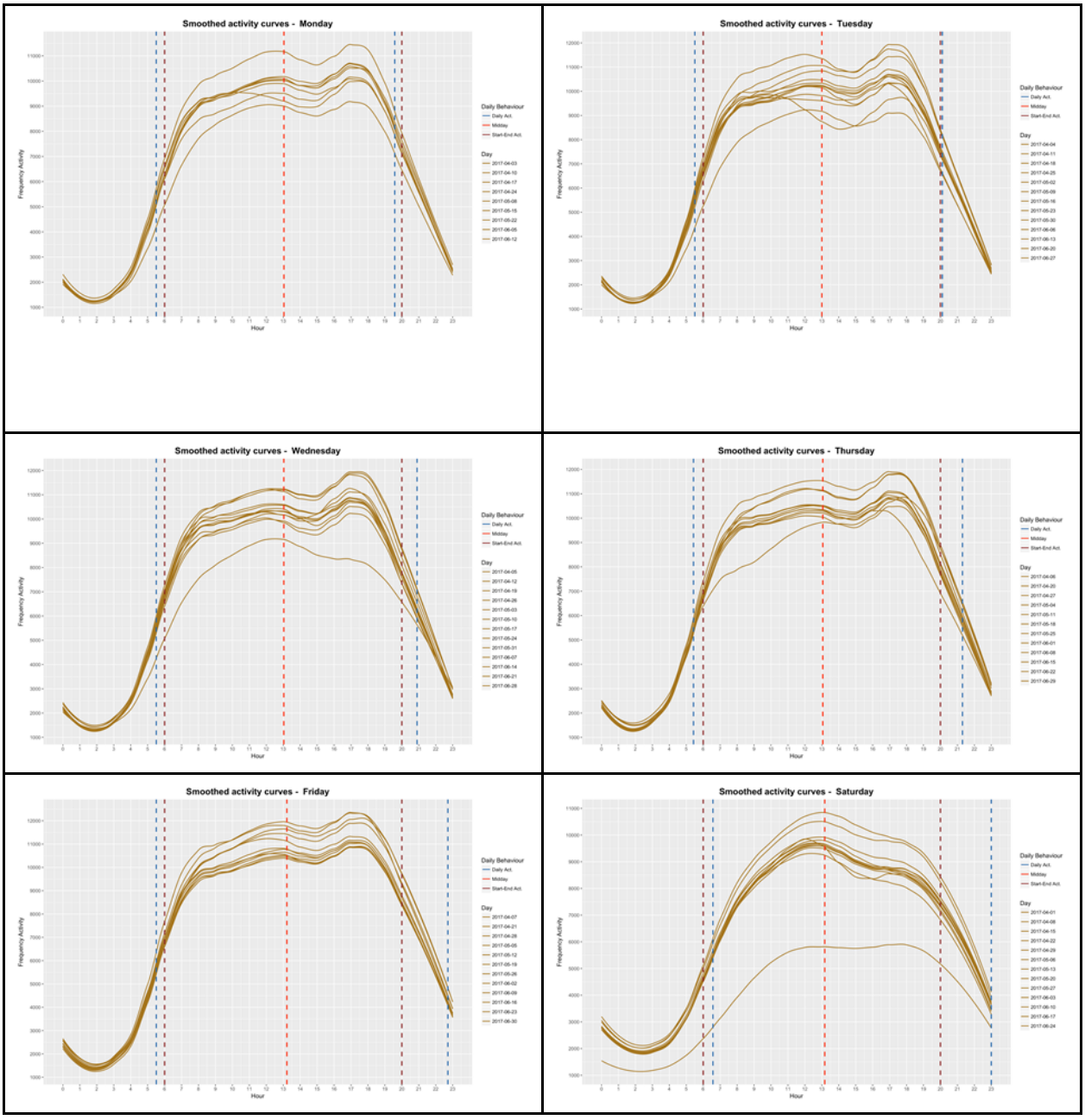
The indicators are calculated using call- and data-related events, such as the number of connections between a device and the base station.

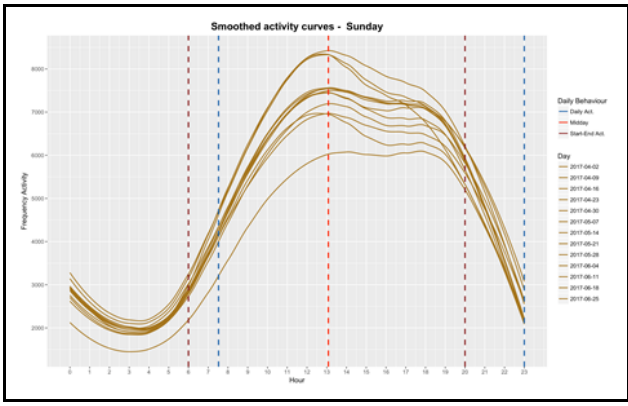
The first indicator proposed groups municipalities according to their daily call and data activity. It will serve to identify clusters of municipalities with similar daily behaviour. The results will be supplemented with socioeconomic information and administrative records, in order to generate analysis value.

The clusters will be generated using as input information on daily and nightly activity, and the duration of the day as determined by call and data activity. The clusters will be generated using the fuzzy clustering method. The results will be visualized on maps in order to facilitate their use.

The results generated to date include smoothed activity curves for each day. Figure 4 shows the results for Bogotá.

Figure 4: Smothed activity curves based on CDR, Bogotá





BD17: CLASSIFICATION OF AREAS AS RESIDENTIAL OR NON-RESIDENTIAL

The objective of the second proposed indicator is to classify areas as residential or non-residential.

For each city or group of regions, the fuzzy clustering method will be applied to all base stations (week-weekends), in order to determine whether they should be grouped and subsequently classified as residential or non-residential based on their level of CDR activity and additional information on ground use and other administrative records, such as energy consumption.

This indicator is being constructed.

BD18: CALCULATION OF ACTIVITY DENSITIES DISAGGREGATED BY AREA

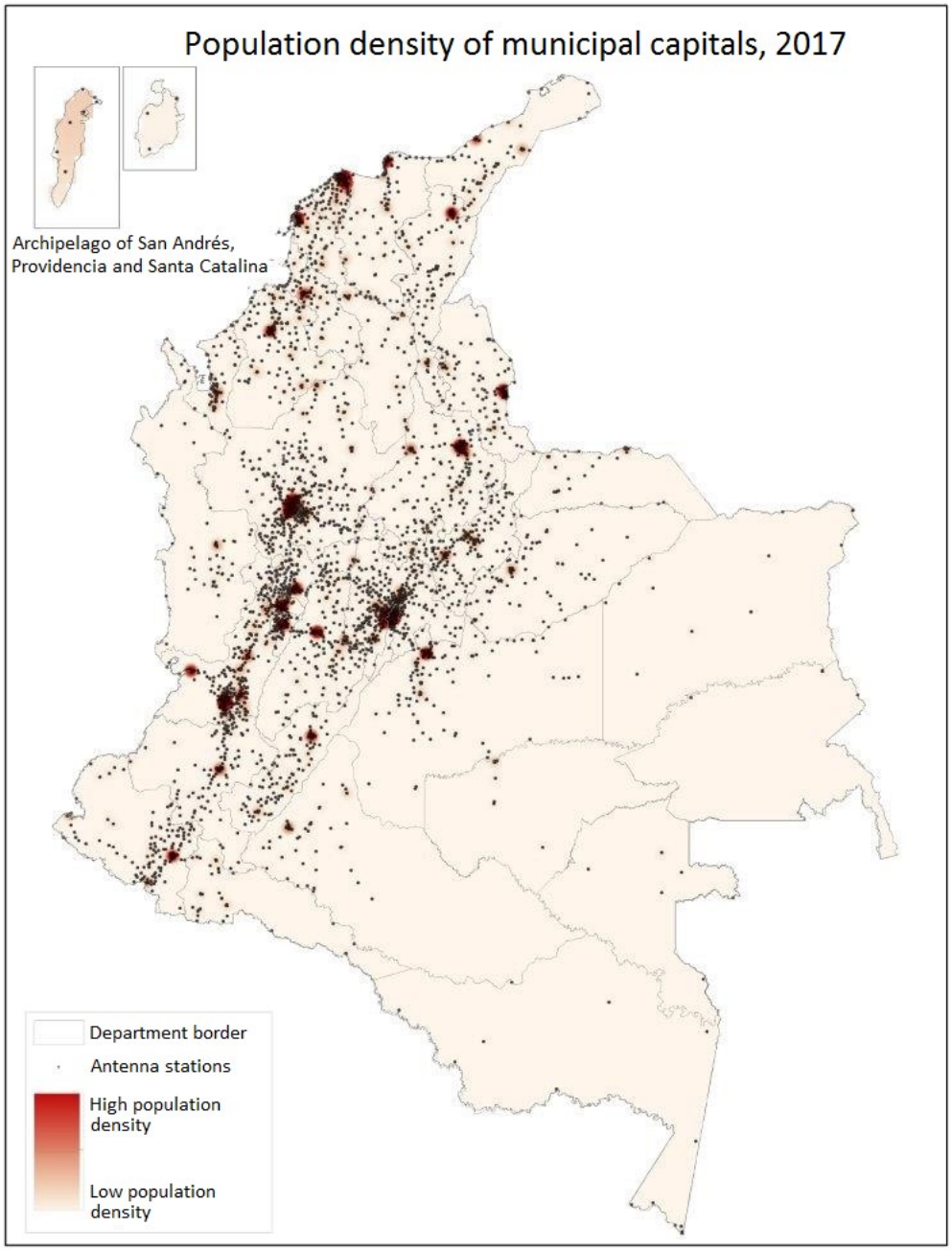
The aim of the third indicator is to provide additional information for the population and housing census, by calculating activity densities for the areas previously classified as residential or non-residential.

This indicator is being constructed.

BD19: DEPLOYMENT OF MOBILE NETWORK INFRASTRUCTURE BY POPULATION DENSITY

The fourth indicator shows the mobile network infrastructure deployed in Colombia by the network and telecommunications service provider Comunicación Celular S.A. (Comcel). The analysis is based on the relationship between demographic information (in terms of population density for each of Colombia's municipalities) and Comcel network (base stations) information at municipal level.

Map 4: Population density and network antenna density



6. RECOMMENDATIONS

Formulating recommendations for projects of this type, which involve the analysis of large volumes of data, can at times be more a matter of logic or common sense than a technical issue. No need to be a data scientist, engineer or economist to interpret a set of results with a view to making suggestions and/or comments for the future development of such efforts.

In view of the above, the Colombian interdisciplinary team put together to develop and carry out the Big Data for Measuring the Information Society pilot project, spearheaded by ITU and carried out in Colombia by MinTEC, DANE and the network and telecommunications service provider Comunicación Celular S.A. (Comcel), has drawn up a series of recommendations for the more than 200 economies worldwide seeking to develop projects for the analysis of large volumes of data supplied by telecommunications service providers.

The recommendations provided aim to contribute to the search for new sources of data for measuring ICT indicators.

I. Start by calculating ICT indicators on a small scale, i.e. using population and/or ICT infrastructure information at municipal (town) level, in order to compare data and obtain reliable results from the outset, and in this way broaden the data set to be analysed to the point where the analyses are being conducted at national level.
II. Although the raw data may comprise millions of records, it is recommended to start by evaluating sample data that are significant in terms of completeness of the anticipated characters per field, in order to avoid reprocessing at the indicator calculation stage.
III. Whenever possible, obtain primary or secondary information via administrative records, censuses, surveys or other sources of information that can be analysed against the results obtained by calculating big data indicators.
IV. Establish an effective channel of communication (such as a digital committee) between the various stakeholders involved in the development of projects of this kind, given that the speed at which results and indicators are generated calls for almost real-time validation of the results if there is to be no delay in the delivery of outcomes.
V. Conclude public-private administrative agreements between national statistics institutes, ICT ministries, ICT regulators, ICT superintendencies and/or other sectoral public organizations, on the one hand, and network and telecommunications service providers, on the other, for the management of the information produced by projects of this kind, in order to enable the coordinated publication of results, without undermining the commercial and growth strategies of telecommunications service providers.
VI. Opting to work with summary tables of the raw data supplied by the telecommunications service provider could be an excellent choice in terms of cost-benefit. It is nevertheless suggested that at least three desktop trials be conducted of the results of the data processing and calculation of big data indicators, in order to identify where summary table bundles need fine-tuning and thereby minimize loss of information and of raw data patterns.
VII. Colombia has a series of administrative acts that together constitute a robust system of ICT sector information, whereby each network and telecommunications service provider is obliged periodically to report on the telecommunication services it provides. However, as demand grows for sector information for public policy and/or decision-making purposes, telecommunications service providers are being called on to provide special information on a daily basis; this can in some cases exhaust the administrative capacity of a provider's customer service offices or its information-reporting capacity. Using big data to generate ICT indicators is a modern and alternative approach. Thus, an investigation or study of the national impact of the step-by-step

dismantling of the reporting system would be a first step towards adopting the use of big data nationwide as a means of searching for new sources of data and of calculating ICT indicators.

VIII. In order to ensure that data transfers between the telecommunications service provider and the Big Data project data centre are complete, it is suggested that the telecommunications service provider use an algorithm to count the records for each of the files to be transferred, thereby informing the project data centre of the number of records the file is expected to contain; indeed, because of the size of files, records can be lost during transmission.

IX. If no dedicated network connection had been configured between the telecommunications service provider and the project data centre for the transmission of data, it is recommended to run file transfer processes between 2100 hours and 0600 hours of the following day, in order to avoid congestion on the LAN network for Internet service users.

7. CONCLUSIONS

In the pilot project, a number of indicators related to ICT were proposed and calculated. There are two main possibilities to assess the value of these indicators:

- On the national level, whether the indicators provide new insight into the ICT situation;
- On the international level, to compare the ICT levels of different countries.

As the country reports are generated after the results have been received, the comparison between countries is presented in the Project Final Report. It is only then that the comparison between countries and the international value of the indicators can be assessed. International comparability can also provide additional insight into the value of the indicators within each country.

One of the objectives of the project was to test the process of generating new indicators based on big data and identify the challenges and possibilities. This information is also valuable as the challenges are often similar in different countries and the results serve as a good insight for other countries wishing to conduct such experiments.

In general, the challenges can be divided into two categories:

- administrative and legal
- technical and methodological.

I. The development of projects that involve access to information in the CDRs of network and telecommunications service providers will always be a huge challenge for national institutions looking for new sources of data for the calculation of ICT indicators, as the providers hold that their infrastructure and business strategies may be undermined as a result. Projects of this kind must therefore be accompanied by confidentiality agreements between all the parties for the handling of the raw data, but not of the bundled data produced after the indicators have been calculated.

II. At the start of projects of this kind, the word most often used by the participating national entities tends to be “perseverance”. Indeed, it is a tiresome undertaking to attract the attention of executives of all the parties involved in the project’s implementation, with obtaining the stakeholders’ commitment being the biggest challenge. The initial project documents must therefore be complete, in order to foster confidence on first reading, ensuring they spark interest, support and a genuine commitment during the project’s lifespan on the part of the parties concerned.

III. The development of confidentiality agreements for projects to analyse large volumes of data supplied by network and telecommunications service providers is one of the most significant challenges we face. It

involves not just drawing up a model confidentiality agreement as a starting point, but also establishing an environment for effective review of the agreement by the technical and legal units of the participating entities. This stage involves much reworking of the agreement, and this has a direct impact on access to data and compliance with the project time-line.

IV. Communication between the technical teams of the participating parties is crucial, especially during the big data indicator calculation phase. Indeed, the validation of the first results obtained by each of the relevant players will spawn confidence among the parties, and ultimately that confidence will ripple outwards until it reaches the general population. The public expects that the indicators generated using big data will provide a real analysis of a specific situation, positioning big data as a new source of data for the generation of ICT indicators.

V. The use of summary tables of raw data to calculate big data indicators is beneficial in terms of availability of data centre computation units and indicator processing and calculation time. Such tables can nevertheless hide fundamental information on the behaviour of the provider's data, and the result of the indicator calculation therefore does not always meet the statistical quality requirements of the technical team participating in the project.

VI. The formulation of special considerations to make up for the absence of a field required to process and calculate a big data indicator using an incomplete record can sometimes be a subject for discussion by the technical team. However, the technical team must sometimes decide to consider the attribution of a variable, in order to generate the calculation of an indicator, and subsequently to validate the statistical consistency of the calculation result, the aim being to adopt it as proxy and simply discard it, and thereby generate the technical requirement for the compilation of the missing field.

VII. One of the most challenging stages for big data projects of this kind may be visualization of the results. After countless administrative, legal and technical efforts, the time comes when the project results must be presented to senior management and the public; the latter has relatively high expectations in terms of accessing the results easily and intuitively. The objective to be pursued when visualizing the results is therefore simplicity, despite the inevitable pressure to use complex tools for that purpose.