IDEP | ECA

WORKING PAPERS

WHICH BIG DATA FOR AFRICA ?

By Makane Faye, Chief (Rtd), Knowledge Services Section, United Nations Economic Commission for Africa

DISCLAIMER

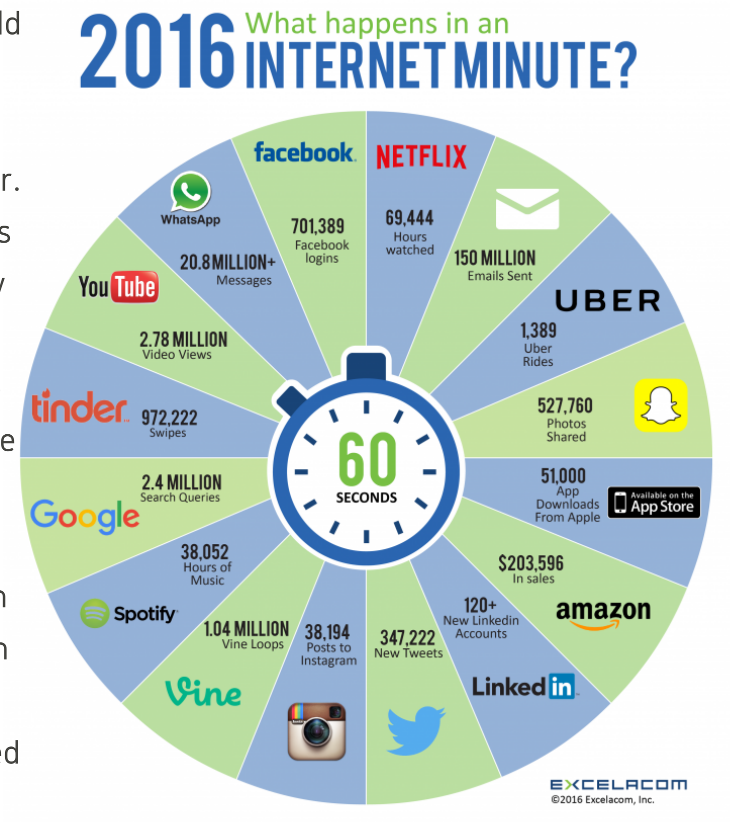
This is a working paper, and hence it represents a reflection in progress intended to stimulate debate debates and moving ideas forward. The views and opinions expressed in it are those of the authors and is the product of professional research work. It is not necessarily meant to reflect the official policy, position or opinions of IDEP, neither its Member countries nor its staff members. Any errors are the fault of the authors.

Copies of papers can be requested from IDEP Library by writing to: Knowledge Management and Library Division, African Institute of Economic Development and Planning (IDEP), Rue du 18 Juin, Dakar, Senegal. Please request papers by number and title.

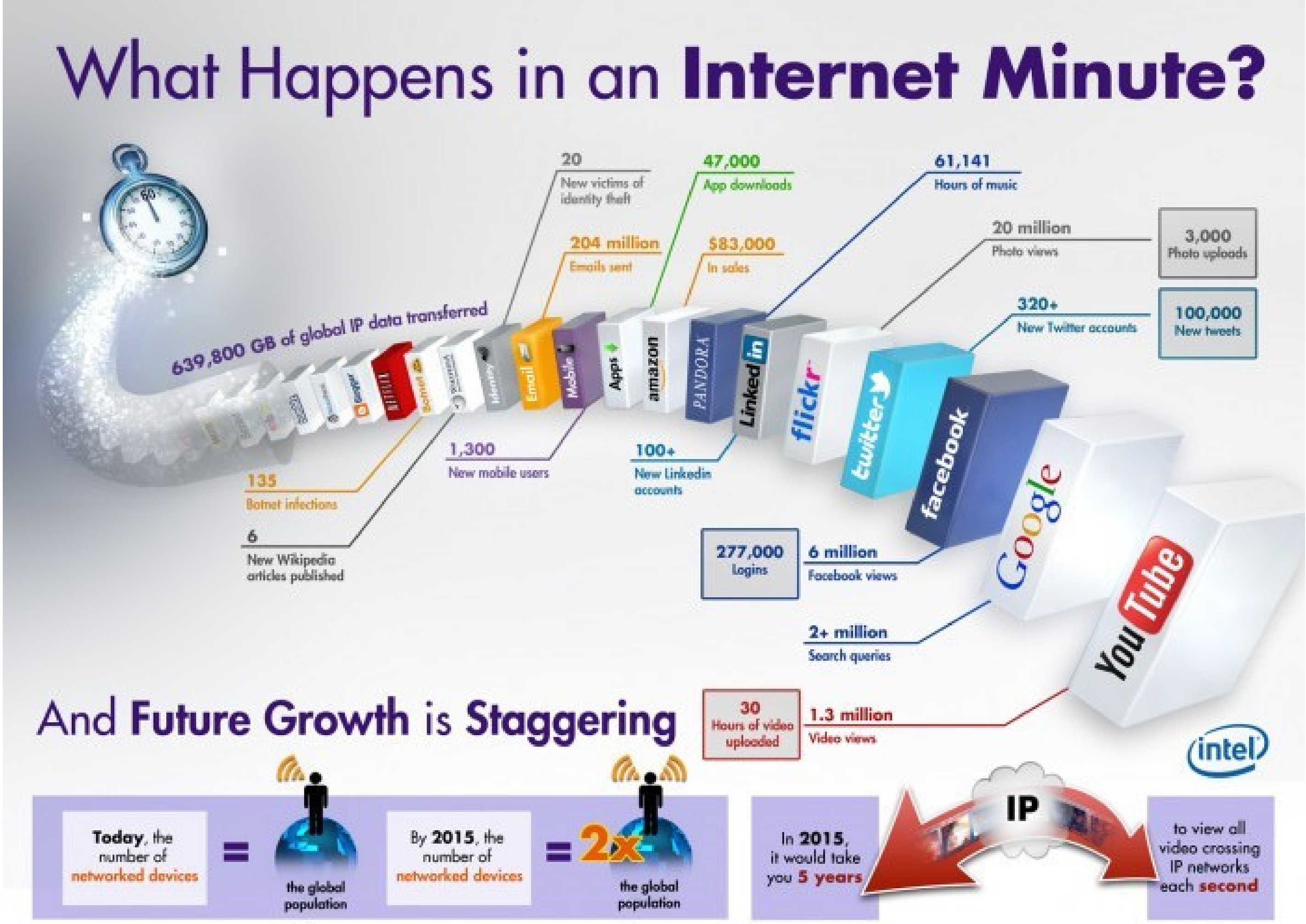
Permission to reprint in whole or in part is hereby granted, provided that the African Institute for Economic Development and Planning is properly cited.

INTRODUCTION

The amount of digital information in the world is growing exponentially, doubling approximately every two years. With the Internet-of-Things, this will move even faster. In 2012, there were less than 9 billion devices connected to the internet. In 2020, some say this number will be more than 50 billion. In 2025, McKinsey expects this number to have increased to 1 trillion. All these devices will be connected; all will be sending data to the cloud. Buckminster Fuller indicated that in 2015, mankind produced as much information as was created in all previous years of human civilization. Buckminster Fuller created the "Knowledge Doubling Curve" where he noticed that until the year 1900, human knowledge doubled approximately every century and that by the end of World War II knowledge was doubling every 25 years. According to IBM in the "Toxic Terabyte" publication, the build out of the "internet of things" will lead to the doubling of knowledge every 12 hours.



WHICH BIG DATA FOR AFRICA ?



According to IBM 2.5 quintillion bytes of data are created every day.

The conceptual note of the IDEP Seminar will guide us throughout our presentation.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

WHICH BIG DATA FOR AFRICA ?

A. Big Data definition (UNECLAC 2016)

Big data is the term used to describe the enormous amount of digital information that is generated when people go about their daily activities, including working, shopping, talking, texting, internet surfing, traveling. It also includes information about natural phenomena generated and/or transmitted by machines, including satellites. The main characteristics of these data that qualify them as “Big data” are that they are large in volume, they are generated very frequently and they cover a multitude of issues from a great number of sources.

B. Big Data Classification

There are several classification attempts grouping the actors/users/systems into communities according to activities and interests

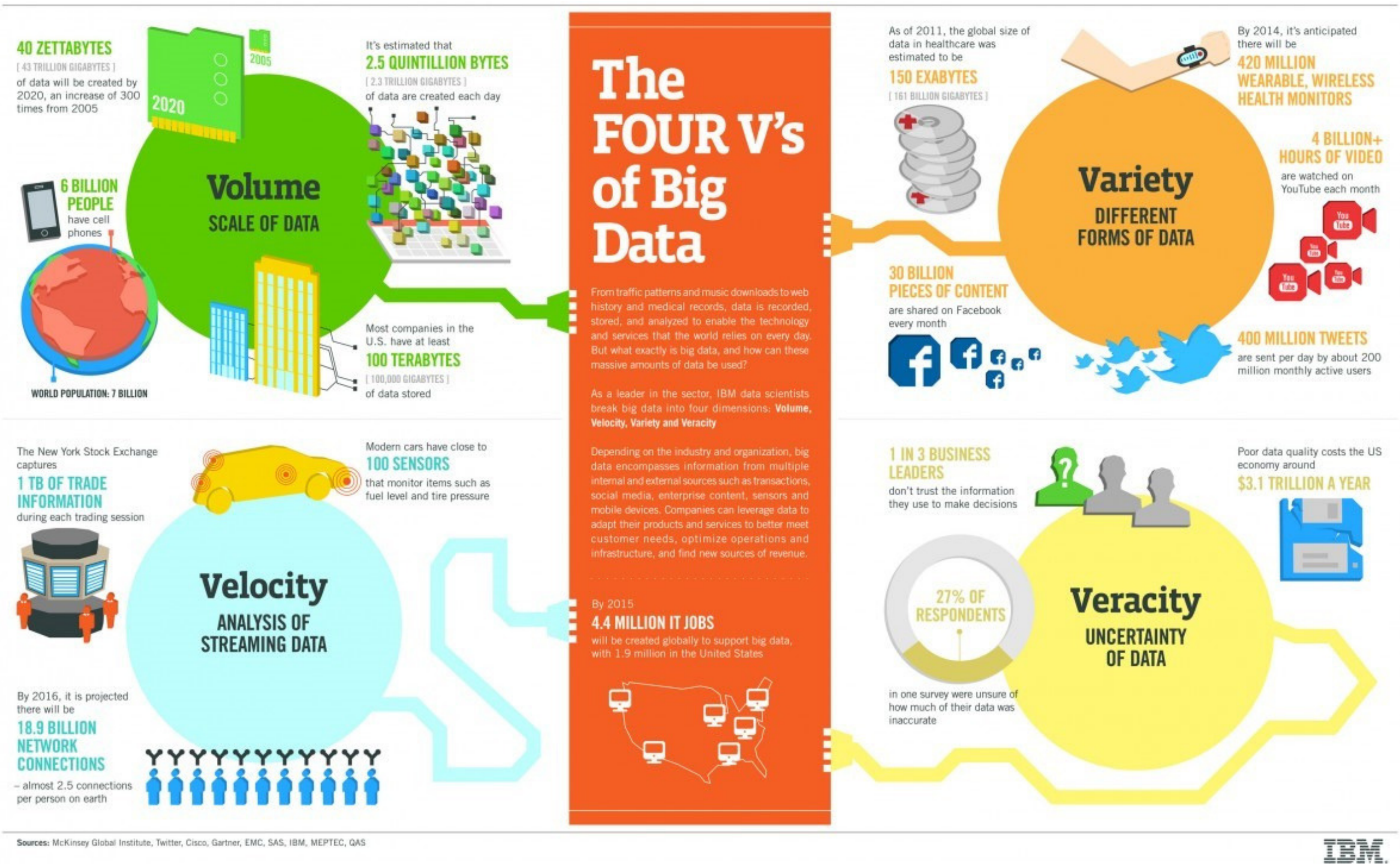
B1. UNECE (Big Data Task Team 2013) put them in 3 main categories:

1. **Social Networks (human-sourced information)**
2. **Traditional Business systems (process-mediated data)**
3. **Internet of Things (machine-generated data)**

According to UNECE, several authors characterize Big Data using the Four Vs definition” that points to the four characteristics of Big Data, namely Volume (the amount of data), Variety (different types of data and sources), Velocity (data on motion) and Veracity (Data uncertainty).

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

WHICH BIG DATA FOR AFRICA ?


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
The FOUR V's of Big Data
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
Variety
DIFFERENT FORMS OF DATA
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
IBM

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

B2. UNECA (2016) Data ecosystems in Africa classification into 9 national data ecosystems :

1. **National statistical systems**
2. **Private-sector data communities**
3. **Civil society data communities**
4. **Academic or scientific data communities**
5. **Open data communities**
6. **Big data communities Citizen-based data communities – Within this community, UNECA refers to definitions from McKinsey, Doug Laney and UN Global Pulse.**

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

C. Big Data Opportunities

Big data has been harnessed worldwide mostly by the private sector, and now governments in the region need to get on board and take full advantage of the opportunities. According to UNECA's Africa Data Revolution Report 2016, African countries with significant private sectors such as Egypt, Kenya, Nigeria, Senegal and South Africa are witnessing the collection, analysis and use of both social media and business data on a considerable scale.

The examples identified by UNECA, coupled with other global experiences, suggest that applications of big data in Africa could be transformative for the achievement of sustainable development, especially in the areas of poverty, agriculture, health, energy, education, innovation and infrastructure, and climate change. Big data can help unlock and generate new insights from the vast quantities of data that are already held by governments, as well as data held by citizen groups and other stakeholders. Integrating data from multiple sources can enable big data to fill in key gaps within official statistics, helping to expand the conception of who and what counts and is counted. Moreover, the early warning, real-time awareness and feedback qualities of big data can help improve programme, policy and project implementation. From the Conference of European Statisticians in 2012, UNECE has been promoting the modernization of statistical production and studying how best to integrate data from multiple sources, including statistical surveys, administrative records, geo-spatial information, data generated by communities and citizens, etc. The report of the United Nations Economic and Social Council, 2013, urges statistical offices at the national, regional and international levels to get fully involved in the processing of Big Data for development.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

D. Confidentiality, Access and Governance

Confidentiality and Privacy issues: In accessing and processing Big Data, what assurances exist on the protection of confidentiality? There are risks regarding privacy and confidentiality (UNECE). The algorithms can lead to wrong decisions. They can wrongly label and stigmatize citizens and divide them into groups, where they do not belong.

For UNECLAC, as the use of big data applications becomes widespread, the negative impact on privacy will certainly increase. Experience has shown that when data are collected and stored, they are eventually reused for purposes undreamt of by the person or company that generated them, or even by some of the big companies that manage them. The anonymization techniques used for small samples are of limited use here, as the innumerable links to and from each piece of data will make its origin easily identifiable by someone with the right technology.

Governance issue: What is the impact on the organization of a (NSO) when Big Data become an important source of data? Compounded by a Methodological one: What is the impact of using Big Data (in combination with, or as a substitute for statistical data) on the consolidated methods of data collection, processing and dissemination?

Transparency issue: We need transparency and access to information. Are Big Data accessible to Government institutions, to researchers and other stakeholders, and under what conditions?

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

E. Follow up of major national, regional and international development agendas

Big Data can pave the way for almost real time data (UNECE). In 2010, when world leaders where assessing the progress towards the Millennium Development Goals, most data was from 2005, which was not satisfactory. UNECE believes that with multi-source statistics and Big Data, when member States will meet to assess progress towards the SDGs in 2030, they should have data that is less than one year old, and it may be possible to have even data from the day before! Moreover when dealing with the development agendas, there are huge possibilities in combining Big Data with “impressive computer power and Machine learning”. This can ensure better predictions and policies and can help to relieve the strain on scarce public budgets.

F. The Technology

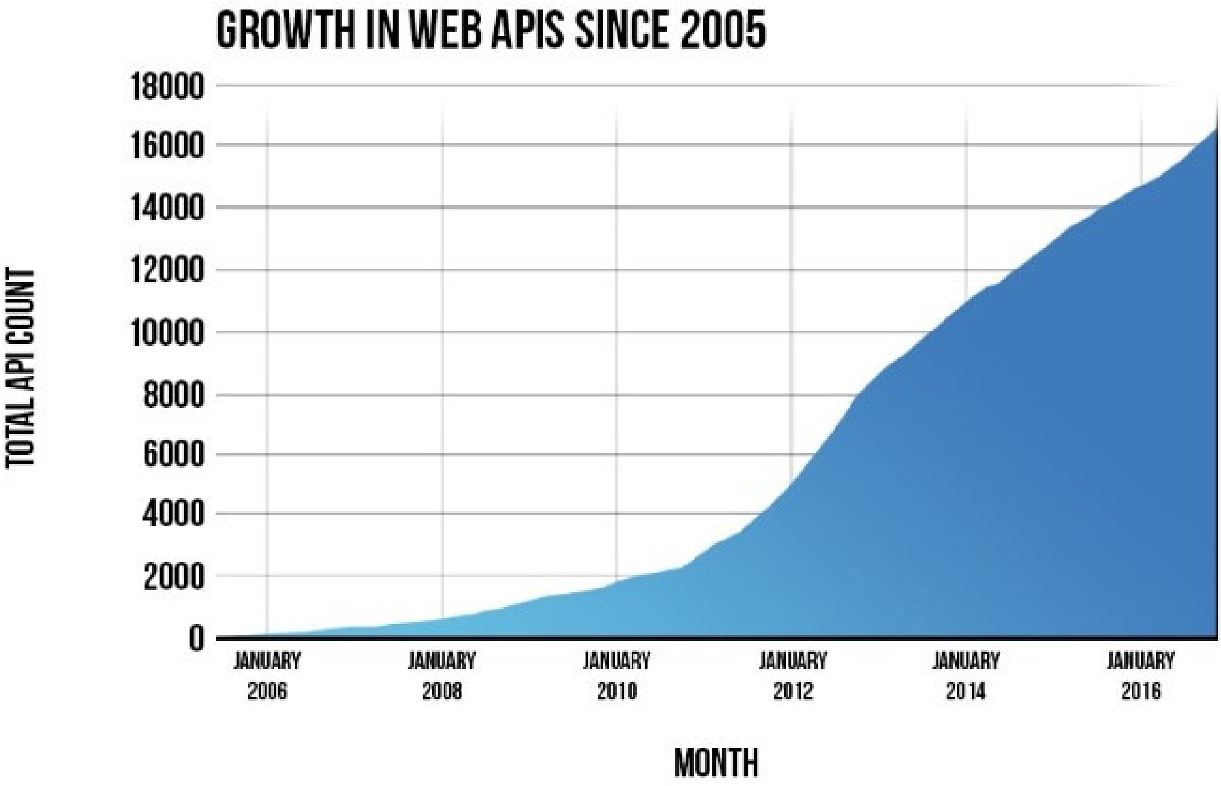
The United Nations Economic and Social Council (2013b) justifies a careful consideration into big data analytics by illustrating how inadequate and inappropriate traditional data management tools are in analysing the volume of data that is now available. It further indicates that it would be unrealistic and expensive to attempt to tailor traditional infrastructure to process big data. These issues call for new means of processing and analysing data.

Martin Hilbert argues that the prerequisite for making big data analytics work for development is ”a solid technological (hardware) infrastructure, generic (software) services and human capacities and skills. These horizontal layers are the requirement sine qua non …”

WHICH BIG DATA FOR AFRICA ?

For UNECLAC, the management and efficient use of big data require new tools for data capture and transmission, in addition to those involved in data analysis. These kinds of tools deserve special attention: the application programming interfaces (APIs), which are a set of software modules, tools, and protocols that enable two or more platforms, systems and most commonly, applications to communicate with each other and initiate tasks or processes. The supply of APIs is growing, with Programmable Web reporting that the number rose from 1 in June 2005 to 16,590 at the end of 2016.

ProgrammableWeb



© 2018 by the African Institute for Economic Development and Planning. All rights reserved

WHICH BIG DATA FOR AFRICA ?

Looking of the nature and sources of Big Data, ECOSOC, 2013b, p.2 indicates that "Data are no longer centralized, highly structured and easily manageable, but are highly distributed, loosely structured (if structured at all), and increasingly large in volume". From this we assume the need for new tools; and for UNECE a paradigm shift is required in Information Technology in order to start using Big Data. UNECE further indicates that Access to Big Data often has a financial cost sometimes considerable. According to research undertaken, most of the APIs are provided free at the beginning for small activities but have cost when it comes to serious enterprise implementation. Hence developers and users should know that everything is not free in the data acquisition, processing and visualization world.

F1. Investing in Big Data and analytics

A recent McKinsey and Company report suggested that companies that invest in big data and analytics consistently outperform their peers in both productivity and revenue. But where to start?

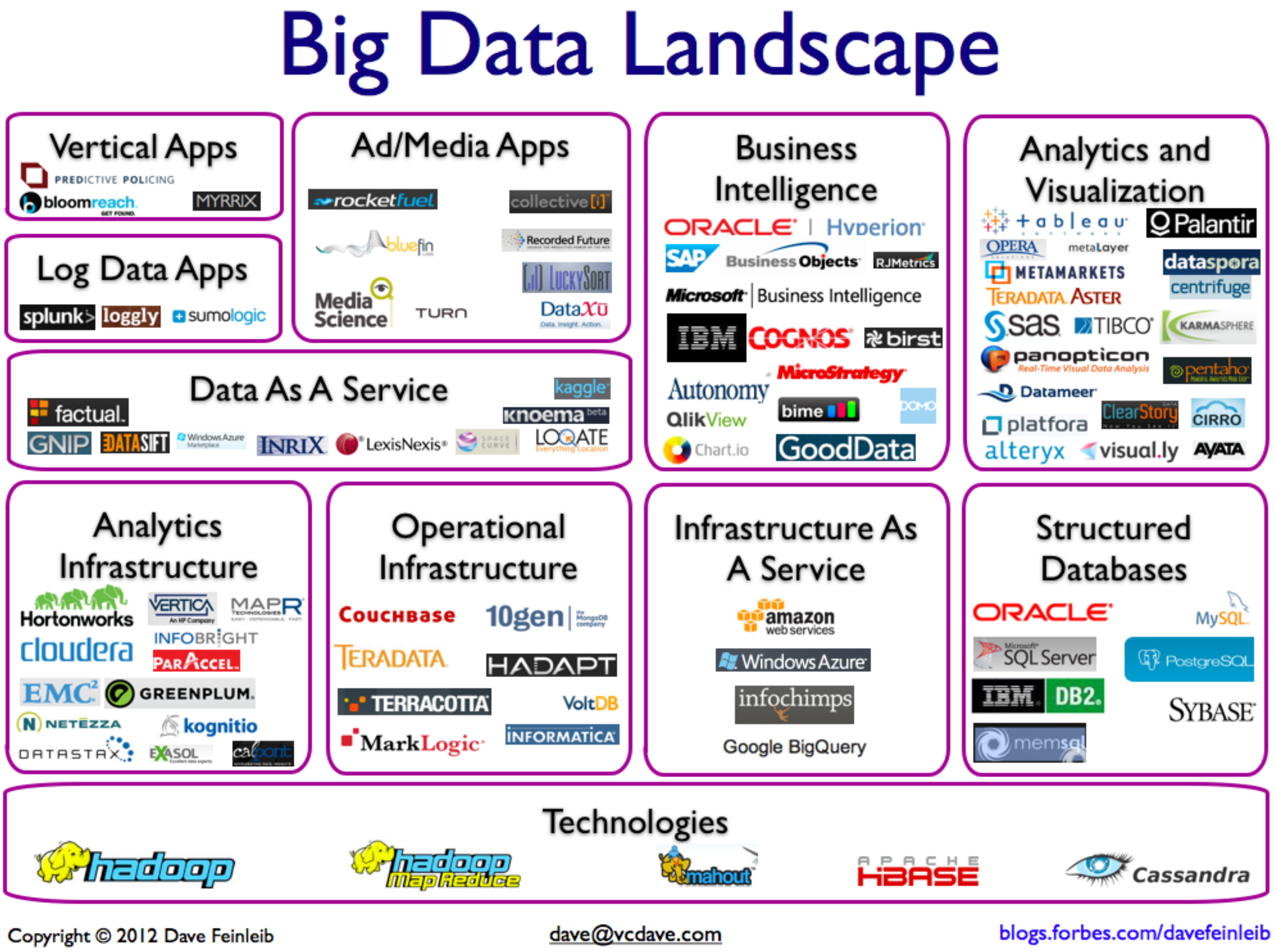
Altitude Digital underlines below the steps to take when making the leap from antiquated databases of old to the modern big data platform.

1. Start small (and free). Hadoop is the open-source software framework of choice for many in the big data game. It's built to scale, and can run on single servers to thousands of machines, and it is designed to handle failures at the application level rather than at the hardware level.

2. Get familiar with the ecosystem. The architecture around the platform can make a big difference in how effective and efficient your business can be. On top of a big data distribution (MapR, Cloudera, and Hortonworks being some of the most used in the space), Hadoop is integrated with a number of tools to make it easier to manage, understand, and use data.

WHICH BIG DATA FOR AFRICA ?

3. It is recommended to have a **provider** who packages these different solutions and makes them easy to experiment with and test. Training is also essential.



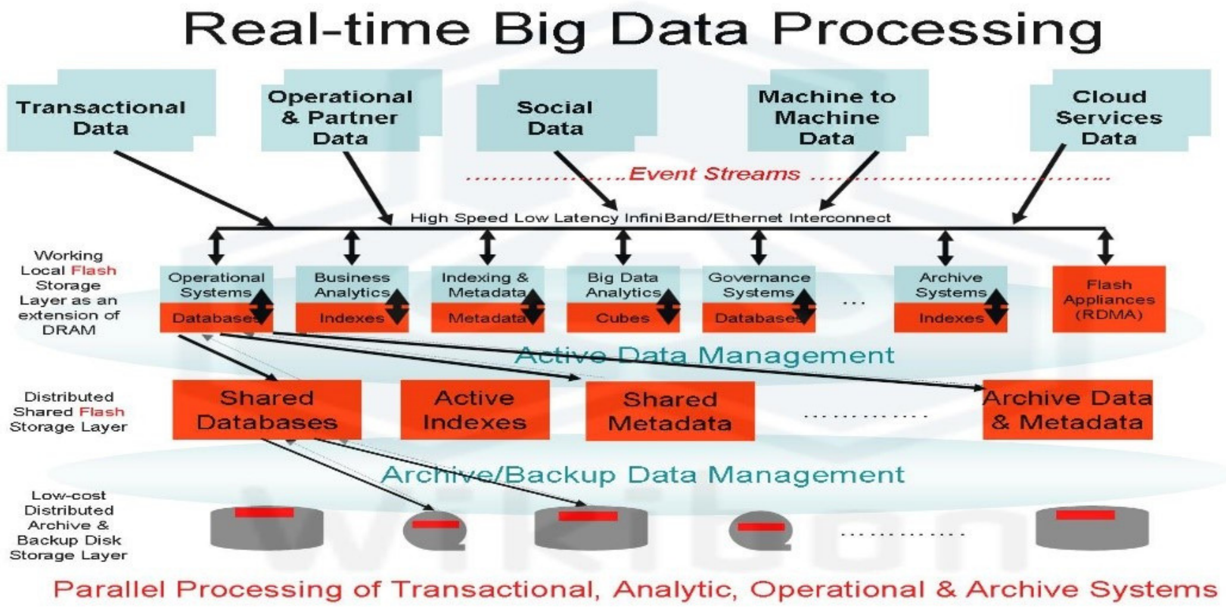
© 2018 by the African Institute for Economic Development and Planning. All rights reserved

F2. According to Forrester, below are the "10 hottest big data technologies":

1. Predictive analytics: software and/or hardware solutions that allow firms to discover, evaluate, optimize, and deploy predictive models by analyzing big data sources to improve business performance or mitigate risk.
2. NoSQL databases: key-value, document, and graph databases.
3. Search and knowledge discovery: tools and technologies to support self-service extraction of information and new insights from large repositories of unstructured and structured data that resides in multiple sources such as file systems, databases, streams, APIs, and other platforms and applications.
4. Stream analytics: software that can filter, aggregate, enrich, and analyze a high throughput of data from multiple disparate live data sources and in any data format.
5. In-memory data fabric: provides low-latency access and processing of large quantities of data by distributing data across the dynamic random access memory (DRAM), Flash, or SSD of a distributed computer system.
6. Distributed file stores: a computer network where data is stored on more than one node, often in a replicated fashion, for redundancy and performance.
7. Data virtualization: a technology that delivers information from various data sources, including big data sources such as Hadoop and distributed data stores in real-time and near-real time.
8. Data integration: tools for data orchestration across solutions such as Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couchbase, Hadoop, and MongoDB.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

9. Data preparation: software that eases the burden of sourcing, shaping, cleansing, and sharing diverse and messy data sets to accelerate datás usefulness for analytics.
10. Data quality: products that conduct data cleansing and enrichment on large, high-velocity data sets, using parallel operations on distributed data stores and databases.


Real-time Big Data Processing
Transactional Data
Operational & Partner Data
Social Data
Machine to Machine Data
Cloud Services Data
Event Streams
High Speed Low Latency InfiniBand/Ethernet Interconnect
Working Local Flash Storage Layer as an extension of DRAM
Operational Systems
Databases
Business Analytics
Indexes
Indexing & Metadata
Metadata
Big Data Analytics
Cubes
Governance Systems
Databases
...
Archive Systems
Indexes
Flash Appliances (RDMA)
Active Data Management
Distributed Shared Flash Storage Layer
Shared Databases
Active Indexes
Shared Metadata
Archive Data & Metadata
Archive/Backup Data Management
Low-cost Distributed Archive & Backup Disk Storage Layer
Parallel Processing of Transactional, Analytic, Operational & Archive Systems

G. Investing in Big Data Infrastructure

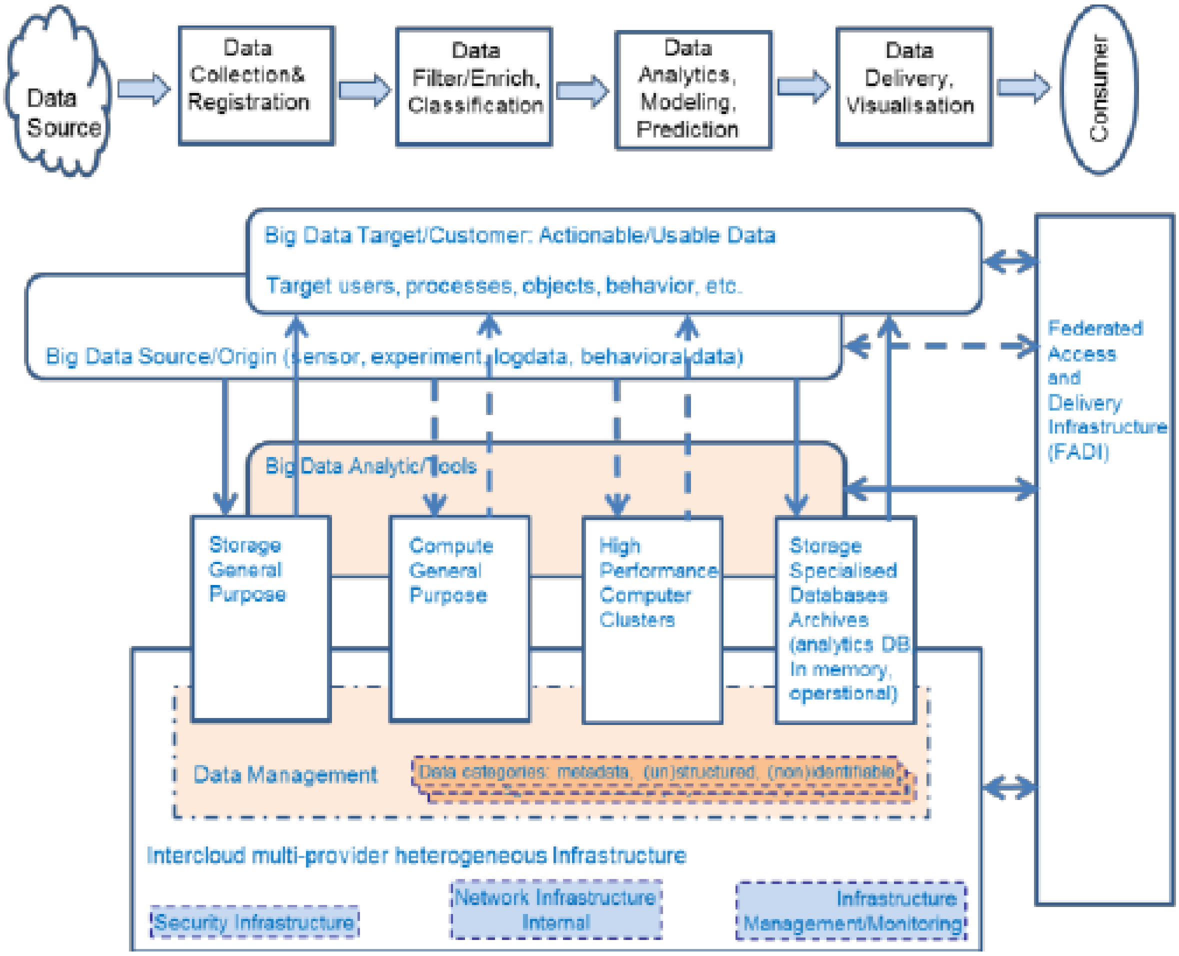
G1. The 4 Big Data Infrastructure Building Blocks

WHICH BIG DATA FOR AFRICA ?

Until recently it was hard for companies to get into big data without making heavy infrastructure investments, but times have changed. Cloud computing in particular has opened up a lot of options for using big data, as it means businesses can tap into big data without having to invest in massive on-site storage and data processing facilities.
In order to get going with big data and turn it into insights and business value, Bernard Marr has identified the following 4 key infrastructure elements as building blocks:

1. Data collection -This is where the data arrives at the company. It includes everything from the sales records, customer database, feedback, social media channels, marketing lists, email archives and any data gleaned from monitoring or measuring aspects of the operations. These data may be in and outside of the company.
2. Data storage - This is where you keep your data once it is gathered from your sources. As the volume of data generated and stored by companies has exploded, sophisticated but accessible systems and tools have been developed to help with this task.
3. Data analysis - This layer is all about turning data into insights. This is where programing languages and platforms come into play. There are three basic steps in this process: a) preparing the data (identifying, cleaning and formatting the data so it is ready for analysis); b) building the analytic model; and c) drawing a conclusion from the insights gained.
4. Data visualization/output - This is how the insights gleaned from analyzing the data are passed on to staff and board members who need them, i.e. the decision makers in the institution. Clear and concise communication is essential, and this output can take the form of brief reports, charts, figures and key recommendations.

WHICH BIG DATA FOR AFRICA ?



Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer
Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.
Big Data Source/Origin (sensor, experiment, logdata, behavioral data)
Federated Access and Delivery Infrastructure (FADI)
Big Data Analytic/Tools
Storage General Purpose
Compute General Purpose
High Performance Computer Clusters
Storage Specialised Databases Archives (analytics DB In memory, operstional)
Data Management
Data categories: metadata, (un)structured, (non)identifiable
Intercloud multi-provider heterogeneous Infrastructure
Security Infrastructure
Network Infrastructure Internal
Infrastructure Management/Monitoring

H. Human Resources

H1. Having the right people and teams may be the big data best practice. There is need for a range of expertise and knowledge.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

H2. According to the Jigsaw Academy, the following 5 skills are essential for Big Data :

1. Skill 1: Programming - Learning how to code is an essential skill in the Big Data analyst's arsenal. Some of the languages needed are Python, R, Java, and C++ among others.
2. Skill 2: Quantitative Skills - Numerical and statistical analysis are core quantitative skills that every good big data analyst needs. This knowledge enables the use of concepts such as neural networks and machine learning.
3. Skill 3: Multiple Technologies – Some staff need to learn multiple technologies that will help them grow as a Big Data analysts. The range of technologies that a good big data analyst must be familiar with is huge. They depend upon the environment staff are working in, which vary based on the requirements of the institution.
4. Skill 4: Understanding of Business & Outcomes - All big data analysts need to have a strong understanding of the business and domain they operate in. Domain expertise can magnify the impact of the big data analyst's insights.
5. Skill 5: Interpretation of Data - It is the one skill that combines both art and science. It requires the precision and sterility of hard science and mathematics but also call for creativity, ingenuity.

I. The UNECE Big Data Project

The UN work in Big Data started in 2012 with a workshop of the European Statistical Offices convened by UNECE, followed by establishment of a UNECE Task Team on Big Data.

WHICH BIG DATA FOR AFRICA ?

To experiment with Big Data, UNECE decided to put in place in 2013 the Sandbox project, a web-based collaborative environment to better understand how to use the power of ”Big Data” to support the production of official statistics. The objectives and outcomes were clearly related to most of the objectives of today's IDEP Seminar.

I.1. The Sandbox project gave participating statistical organizations the opportunity to:

1. Test the feasibility of remote processing of Big Data on a central server
2. Test how existing statistical standards / models / methods can be applied to Big Data
3. Determine which Big Data software tools are most useful for statistical organizations
4. Learn more about the potential uses, advantages and disadvantages of Big Data sets
5. Build an international collaboration community to share ideas and experiences on using Big Data

I.2. Some of the findings of the outcomes of the project, which was concluded in 2015 are as follows :

1. Big Data sources can cover aspects of reality that are not covered by traditional ones. For example, the analysis of Twitter data can give a direct indication of the sentiment of users.
2. Producing statistics based on Big Data would mean accepting different notions of quality.
3. The fact that data are produced in large amounts does not mean they are immediately and easily available for producing statistics. “Quality” sources, meaning data that are particularly relevant and clean, are more difficult to access.
4. Publicly accessible data sources are limited in terms of quality and therefore a significant amount of processing may be required to render them usable for analysis.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

5. Indications on possible ways to enhance the use of IT tools in statistical organizations to both cope with Big Data and to improve the general efficiency of data treatment. In particular, although Big Data technologies were conceived specifically for handling data of significant size, they can also be used effectively to process data of “average” size in a more efficient way than “traditional” tools. Processing the UNECE ComTrade data in the Sandbox provided significant advantages in terms of processing time compared to the relational database system currently used.
6. Learning new IT tools with this level of specialization is a difficult task, especially for statistical researchers, not used to managing complex software platforms.
7. The collaboration model based on the Sandbox environment facilitated sharing of knowledge about tools, methods and solutions, and created a new form of cooperation between statisticians and the IT sector.
8. The availability of a common environment allowed new tools to be tested without having to set up costly and complex IT infrastructures.

J. Conclusion

As the United Nations Global Pulse initiative has argued, big data do not and should not be expected to contain all the answers to human problems. They also have their own limitations and biases, which must be understood and taken into account. Finally, ”real-time information does not replace the quantitative statistical evidence which governments traditionally use for decision-making, but if understood correctly, it can inform where further targeted investigation is necessary, or even inform immediate response if necessary, and thus change outcomes like nothing else can”.

RECOMMENDATIONS

RECOMMENDATION FROM UNECA

According to UNECA, greater collaboration and coordination between data communities can significantly reduce the costs of data collection for use in pursuing the Sustainable Development Goals, Agenda 2063 and long-term national development plans by helping fill gaps in official statistics and enhance data quality, timeliness, relevance, accessibility, dissemination and use.

RECOMMENDATION FROM UNECE

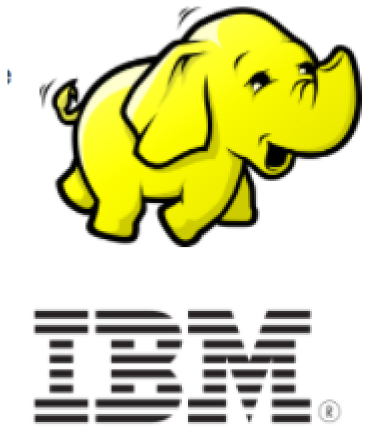
UNECE promotes building of new partnerships – with academia, with other authorities – from schools and hospitals to public utilities, and not least with the private sector, from electricity suppliers to Google, Amazon and Facebook. And finally we need partnerships with citizens and communities in each and every country. Citizens will generate the data. Just as we see crowd-sourcing and crowd-funding we need crowd-data.

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

WHICH BIG DATA FOR AFRICA ?

K. Selected list of Big Data Processing, storage and visualization tools.

Hadoop



Cloudera

DataCleaner
cloudera
mongoDB
rapidminer
talend
TRIFACTA
Quantum

Oracle data mining

ORACLE
TERADATA
optimove
Qubole
kaggle
R

© 2018 by the African Institute for Economic Development and Planning. All rights reserved

WHICH BIG DATA FOR AFRICA ?

bigml

STATWING
MODERNIZING
DATA ANALYSIS

THOUGHTSPOT

DOMO

CARTO

RegEx
Regular Expression

Silk

+ a b | e a u

<xpath>

plotly

python™

chartio

Ideata™

Datawrapper

Google Developers

Blockspring

informatica

magestore
Best Retail Solution for Magento®

pentaho

Stitch

© 2018 by the African Institute for Economic Development and Planning. All rights reserved