FUNDAMENTALS OF BIG DATA

Data analysis & ML

Marcelo Horacio Fortino
MBA PM | PSM I | ITIL & ISO 20000
www.fortinux.com | @HoracioGRC

Copyright

Based on the following works:

SKIENA, Steven S. The Data Science Design Manual. (2017). Switzerland: Springer.

KUBAT, Miroslav. An Introduction to Machine Learning. Second Edition (2017). Switzerland: Springer.

TURKINGTON, Garry. Hadoop Beginner's Guide. (2013). UK: Packt Publishing.

O'REILLY RADAR TEAM. Planning for Big Data. (2012). USA: O'Reilly Media.

Objectives

- **Know Big Data solutions and technologies such as Apache Hadoop.**
- **Acquire knowledge to design business intelligence strategies integrating large data sets and data warehouses.**
- **Develop Machine Learning in-house using Spark MLlib and TensorFlow.**

Contents

- Introduction to Big Data and Analytics.
- Big Data market and trends.
- Big Data definition and history.
- Types of Big Data: structured, unstructured, semi-structured.
- Big Data use cases.
- Best practices for Big Data analytics.
- Hadoop: HDFS & MapReduce, YARN.

Contents

- Big Data processes: ingest, store, process/query, visualize.
- Tools and technologies: Hadoop, Sqoop, Kafka, Mesos, Redis, CouchDB.
- Document stores: MongoDB.
- Column stores: HBase + Cassandra.
- Big Data analytics: Spark, Storm.
- Elastic Stack: Logstash, ElasticSearch and Kibana.
- Machine learning techniques:
- Spark (MLlib, Streaming).
- TensorFlow.

How can data science and big data help my organization?

Marcelo Horacio Fortino | MBA – ITIL & ISO 20000 | www.fortinux.com | @HoracioGRC

Big Data

- **May I extract meaningful insights about the trends, correlations and patterns that exist within big data?**

Big Data

- **In the past only big business could afford to profit from big data.**
- **Walmart, Google, specialized financial traders.**

Big Data

- **Today with Hadoop, commodity Linux hardware and cloud computing, almost everyone can do it.**

There is a data revolution

Marcelo Horacio Fortino | MBA – ITIL & ISO 20000 | www.fortinux.com | @HoracioGRC

Zettabytes

Multiples of bytes						V • T • E
Decimal			Binary			
Value		Metric	Value		IEC	JEDEC
1000	kB	kilobyte	1024	KiB	kibibyte	KB kilobyte
1000^2	MB	megabyte	1024^2	MiB	mebibyte	MB megabyte
1000^3	GB	gigabyte	1024^3	GiB	gibibyte	GB gigabyte
1000^4	TB	terabyte	1024^4	TiB	tebibyte	–
1000^5	PB	petabyte	1024^5	PiB	pebibyte	–
1000^6	EB	exabyte	1024^6	EiB	exbibyte	–
1000^7	ZB	**zettabyte**	1024^7	ZiB	zebibyte	–
1000^8	YB	yottabyte	1024^8	YiB	yobibyte	–
Orders of magnitude of data						

Font: Wikipedia.
https://en.wikipedia.org/wiki/Zettabyte

- **IDC predicts that the Global Datasphere will grow from 33 Zettabytes in 2018 to 175 Zettabytes by 2025.**

Font: The Digitization of the World – From Edge to Core. IDC White Paper. Doc# US44413318. November 2018.

Zettabytes

- **By 2025, every connected person in the world on average will have a digital data engagement over 4,900 times per day – that's about 1 digital interaction every 18 seconds.**

Font: The Digitization of the World – From Edge to Core. IDC White Paper. Doc# US44413318. November 2018.

Zettabytes

- **"The data-driven world will be always on, always tracking, always monitoring, always listening and always watching – because it will be always learning."**

Font: The Digitization of the World – From Edge to Core. IDC White Paper. Doc# US44413318. November 2018.

Big Data Market

- **The global Big Data and business analytics market has grown healthy over the past few years.**
- **$122 billion in global revenue in 2015, and an estimated $189 billion in 2019.**
- **IDC projects that revenue will grow up to $274 billion by 2022.**

Font: PCMag. https://www.pcmag.com/news/the-big-data-market-is-set-to-skyrocket-by-2022. June 2019.

Big Data Market

- **Snowflake to a $69B market cap company.**
- **Palantir reaching a market cap of $22B.**
- **Datadog was a $12B market cap company. 8 months later: $31B.**

Font: Matt Turck. https://mattturck.com/data2020/. September, 2020.

Big Data Trends

- **Operationalization of Big Data.**
- **Fewer unicorns in the data and AI landscape.**
- **Increased alignment between traditional analytics with ML and AI analytics.**

Font: Infoworks.io https://www.infoworks.io/big-data-trends/. October 2019.

Big Data Trends

- **From Hadoop to cloud services to Kubernetes + Snowflake.**
- **Data governance, cataloging, lineage: the increasing importance of data management.**
- **The rise of an AI-specific infrastructure stack (“MLOps”, “AIOps”).**

Font: Mattturck.com. https://mattturck.com/data2020/. September, 2019.

Big Data Trends

- **ETL vs ELT.**
- **Automation of data engineering.**
- **Rise of the data analyst.**
- **Data lakes and data warehouses merging.**
- **Boom time for data science and machine learning platforms (DSML).**
- **GAFAM, Uber, Lyft, etc. have become full-fledged AI companies.**
- **Rise of NLP, a branch of artificial intelligence focused on understanding natural language.**

Font: Mattturck.com. https://mattturck.com/data2020/. September, 2019.

Big Data definition

- **Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.**

Font: Gartner’s Glossary.
https://www.gartner.com/en/information-technology/glossary/big-data

3 V’s of Big Data

- **Volume: Processing of high volumes of low-density, unstructured data. Data of unknown value sometimes, but this might be tens of terabytes or petabytes of data.**

Font: Douglas Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety.

3 V’s of Big Data

- **Velocity: Velocity is the fast rate at which data is received and (perhaps) acted on.**

Font: Douglas Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety.

3 V’s of Big Data

- **Variety: Many types of unstructured and semistructured data types that are available such as text, audio, video, IoT data, etc.**

Font: Douglas Laney. 3D Data Management: Controlling Data Volume, Velocity and Variety.

Anothers V's of Big Data

- **Value: Data has intrinsic value. But it's of no use until that value is discovered.**
- **Veracity: If your data is truthful and how much you can you rely on it.**

Font: What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data.html.

Big Data Types

- **Structured Data: it's already stored in relational databases.**
- **Unstructured Data: examples include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery.**
- **Semi-structured Data: NoSQL documents are considered semi-structured because they contain keywords that can be used to process it.**

Big Data File Formats

- **Benefits of choosing an appropriate file format:**
- **➔ Faster read/write times.**
- **➔ Split across multiple disks.**
- **➔ Support for schema evolution.**
- **➔ Compression support.**

Big Data File Formats

- **Column-oriented data stores are optimized for read-heavy analytical workloads.**
- **Row-based databases are best for write-heavy transactional workloads.**

ORC: https://orc.apache.org/
Parquet: https://parquet.apache.org/documentation/latest/
Avro: https://avro.apache.org/

Big Data File Formats

- **Optimized file formats for Hadoop:**
  - **Apache Optimized Row Columnar (ORC).**
  - **Apache Parquet (column storage).**
  - **Apache Avro (row storage).**

Big Data File Formats

- **Other file formats:**
- → **JSON.**
- → **CSV.**
- → **XML.**

Image: JSON logo: https://www.json.org/json-en.html

Big Data is important

- **The importance of big data analytics has increased along with the variety of unstructured data that can be mined for information: social media content, texts, clickstream data, and sensors from the Internet of Things.**

Font: What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data.html.

Big Data benefits

- **Cost reduction.**
- **Discovering more efficient ways of doing business.**
- **Better decision making.**
- **Create new products and services that customers want and need.**

Font: What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data.html.

History of Big Data

- **Large data sets started in 1960s and '70s with the first data centers and the development of the relational database (SQL language).**

Data Warehouses

- **Inmon coined the term data warehousing promoting the building, usage, and maintenance of data warehouses and related topics.**
- **He wrote the books "Building the Data Warehouse" (1992, with later editions) and "DW 2.0: The Architecture for the Next Generation of Data Warehousing" (2008).**

Font: Wikipedia. https://en.wikipedia.org/wiki/Bill_Inmon

Data Warehouses

- **Today we have the scalability and elasticity of cloud data warehouses:**
- **Amazon Redshift.**
- **Snowflake.**
- **Google BigQuery.**
- **Microsoft Synapse.**

Data Lakes - Warehouses

- **Data lakes are big repositories for raw data, in a variety of formats, that are low-cost, very scalable but don’t support transactions, data quality, etc. (ML)**
- **Data warehouses instead has structured data, transactional capabilities and governance. (BI)**

New paradigm

- **In 2005, the amount of data that was generated through Facebook, Google, and other online services start to become huge.**
- **So in 2006 engineers at Yahoo created Hadoop and launched it as an Apache open source project.**
- **The distributed processing framework made it possible to run big data applications on a clustered platform. NoSQL also began to gain popularity during that time.**

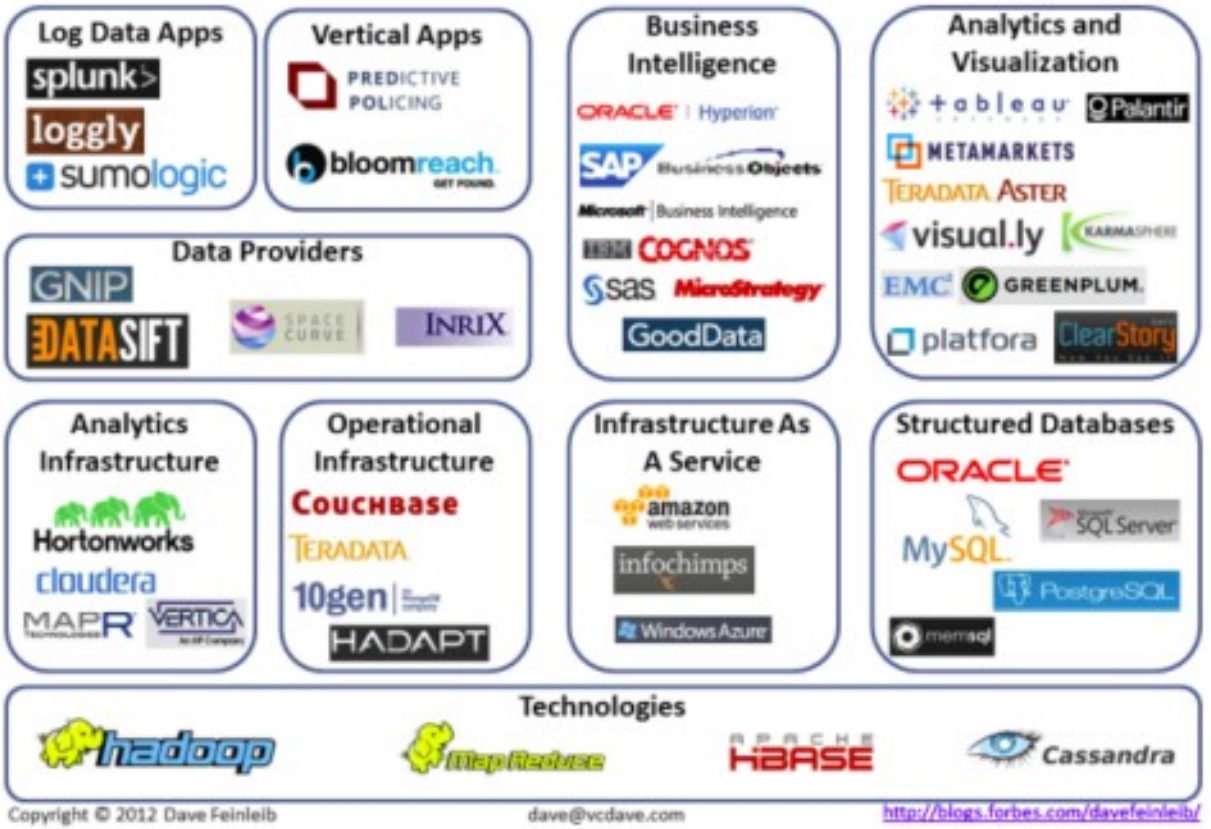
Hadoop and Spark

- **The development of open-source frameworks, such as Hadoop and Spark, was essential for the growth of Big Data because they make easier to work with and cheaper to store.**

Hadoop and cloud platforms

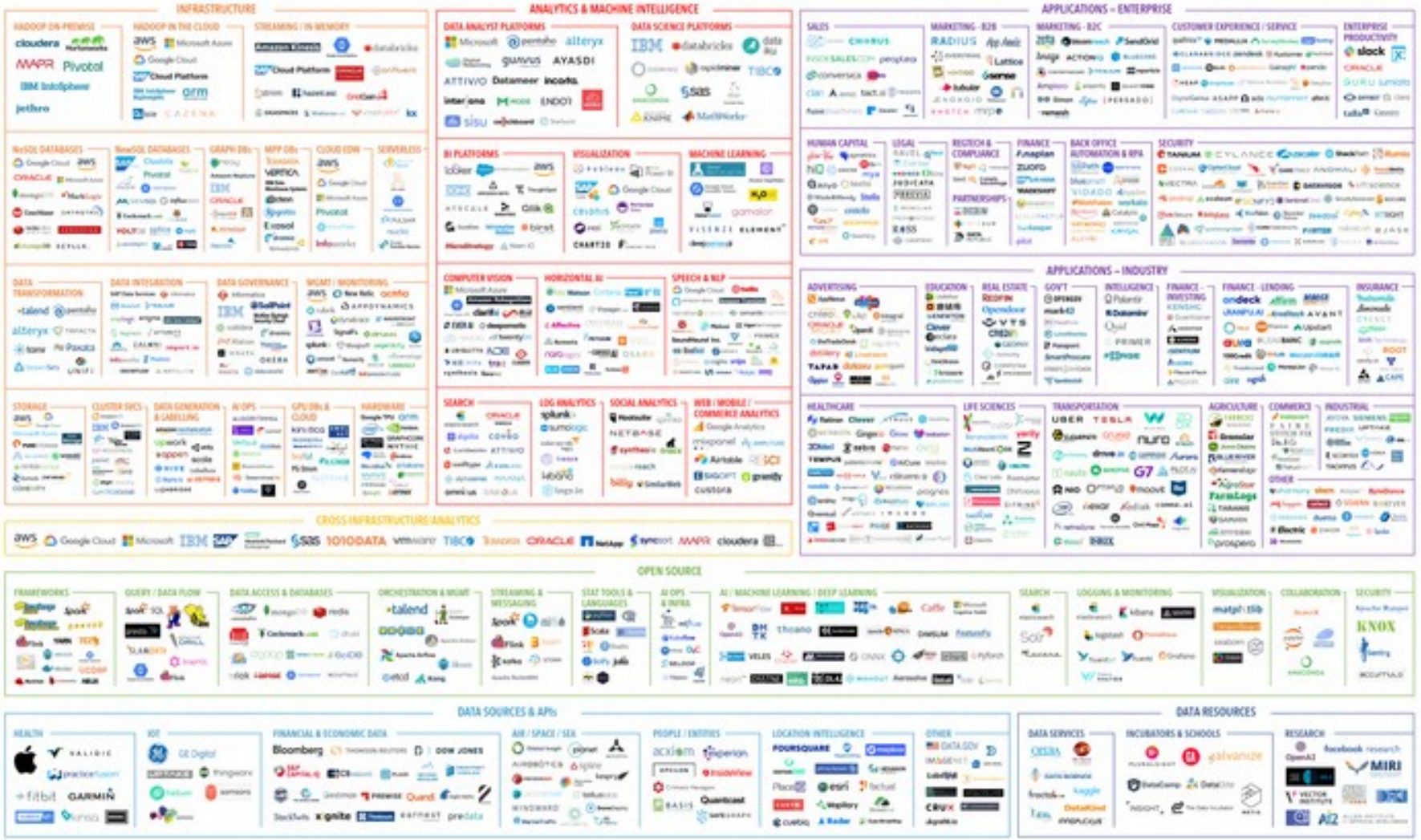
- **Nowadays cloud platform vendors allow almost any business to use a big Data analytics platform setting up Hadoop clusters on demand running only what they need.**

Big Data Landscape



Source: http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/

DATA & AI LANDSCAPE 2019



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Big Data

- Big data use cases

Airline Industry

- **Airlines collect a large volume of data like customer flight preferences, traffic control, baggage handling, aircraft maintenance, flight paths, flight routes, and more.**
- **Big data give insights for optimize operations and better customer service.**

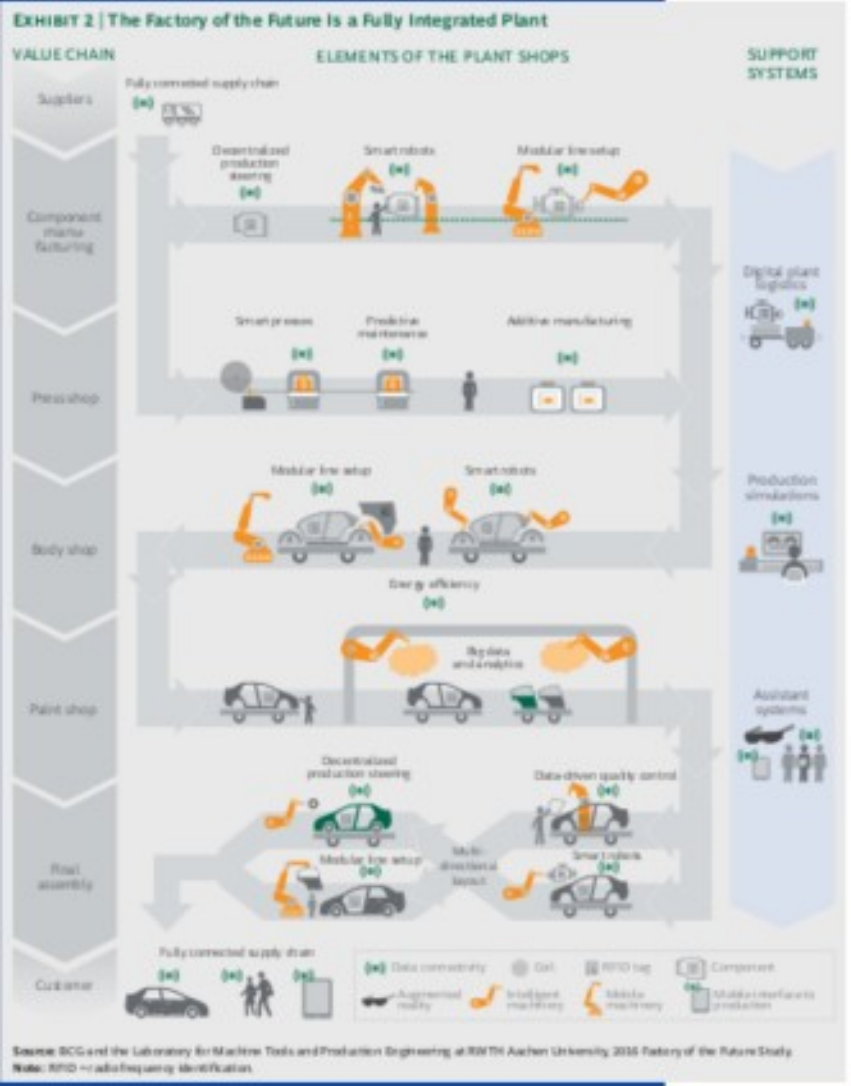
Big Data in Banking

- **The massive amounts of structured and unstructured data that financial institutions can have helps them make better financial decisions.**
- **They also can prevent fraud using big data analytics.**

Big Data in Healthcare

- **Researchers try unprecedented data sharing and cooperation to understand COVID-19 and develop a model for diseases beyond the coronavirus pandemic.**

Font: Big Data and Collaboration Seek to Fight COVID-19. https://www.the-scientist.com/news-opinion/big-data-and-collaboration-seek-to-fight-covid-19-67759



Big Data in Manufacturing

- **Industry 4.0**
- **Big data analytics allows manufacturers to better understand how their value chain works.**
- **It is also used for preventative maintenance of equipment.**

Image: BCG. https://www.bcg.com/publications/2016/leaning-manufacturing-operations-factory-of-future.aspx

Big Data in Science

- ***“Researchers across all disciplines see the newfound ability to link and cross-reference data from diverse sources as improving the accuracy and predictive power of scientific findings and helping to identify future directions of inquiry, thus ultimately providing a novel starting point for empirical investigation.”***

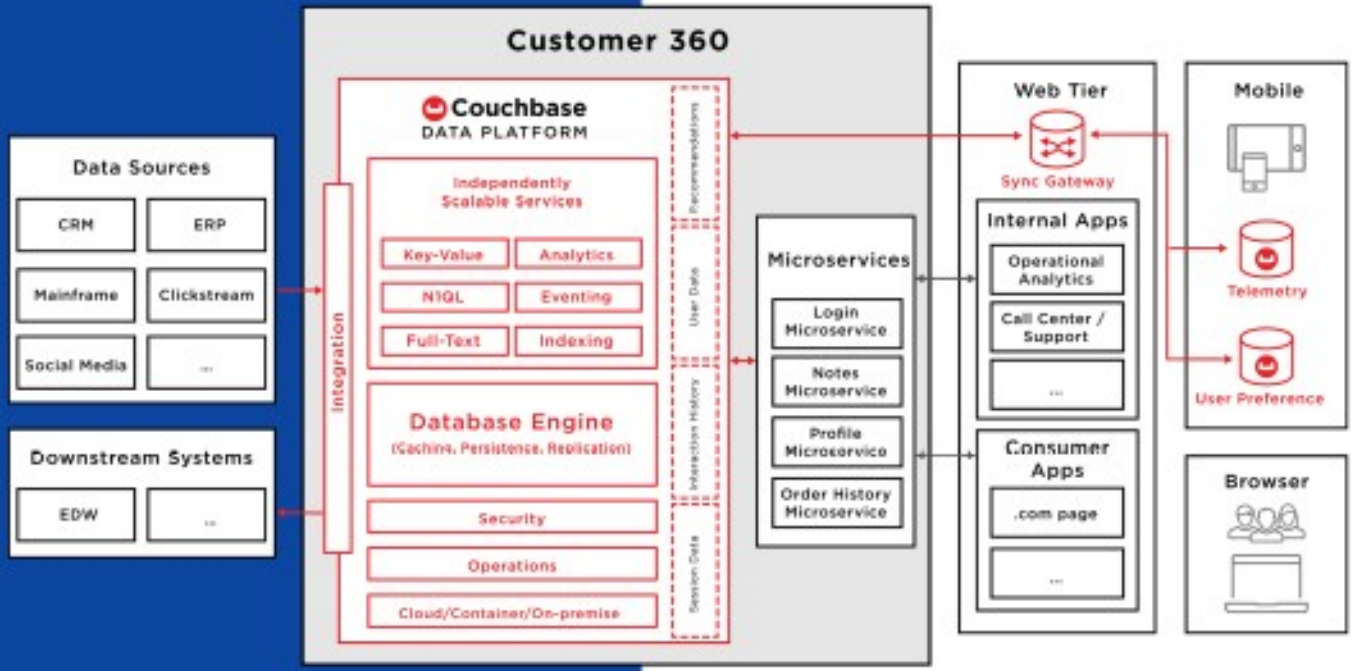
Font: Leonelli, Sabina, "Scientific Research and Big Data", The Stanford Encyclopedia of Philosophy (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/science-big-data/>.

Big Data in Science

- ***"When I started hiring Ph.D. students 15 years ago, they were entirely wet lab," Corcoran says. "Now when we recruit them, the first thing we look for is if they can cope with complex bioinformatic analysis."***
- ***To be a biologist, nowadays, you need to be a statistician, or even a programmer. You need to be able to work with algorithms.***

Font: Phys.org. How big data is changing science. https://phys.org/news/2018-10-big-science.html

Big Data in Retail



- **With big data analytics, retailers can understand customer behavior and preferences, buying habits and demands; predicting trend.**

Image: Couchbase. C360 reference architecture.
https://resources.couchbase.com/sdr/customer-360-view-use-case

Big Data

- Best Practices for Big Data Analytics

Some considerations

- **Big data analytics use data from both internal and external sources.**
- **The data has to be well organized and managed to achieve the best performance.**

Example tools

- **For real-time big data analytics we use a stream processing engine like Spark for data flows through a data store.**
- **For raw data (data lake) to analyze we use the Hadoop Distributed File System.**

Big Data analysis

- **Descriptive analysis – Provides insights on historical data.**
- **Predictive analysis – Provides insights in future data.**
- **Prescriptive analysis – Provides advisable analytics reports for the future.**

Descriptive analysis

- **Data mining : used to sift through data sets in search of patterns and relationships.**

Descriptive analysis

- **What is in your data?**
- **→ Simple metrics, entity lists.**
- **→ Clustering.**
- **→ Segmentation.**
- **→ Dimensional reduction.**

Predictive analysis

- **Predictive analytics: building models to forecast customer behavior.**

Predictive analysis

- **What will the outcome be for a new input?**
- **→ Classification.**
- **→ Regression.**
- **→ Point estimates vs. full distribution.**

Prescriptive analysis

- **What action should we take with this data?**
- **➔ Optimization.**
- **➔ Causal inference.**

Prescriptive analysis

- **Machine learning: programming algorithms to analyze large data sets.**

Prescriptive analysis

- **Deep learning: algorithms that can determine the accuracy of a prediction on their own.**

Big data analytics and business intelligence

- **Business intelligence relies on structured data in a data warehouse and can show what and where an event happened.**

Big data analytics and business intelligence

- **Big data analytics uses both structured and unstructured datasets while explaining why events happened.**
- **It can also predict whether an event will happen again.**

Big Data best practices

- **Align big data with specific business goals.**
- **Ease skills shortage with standards and governance.**
- **Optimize knowledge transfer with a center of excellence.**

Font: What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data.html.

Big Data best practices

- **Top Payoff is aligning unstructured with structured data.**
- **Plan your discovery lab for performance.**
- **Align with the cloud operating model.**

Font: What is Big Data?. Oracle. https://www.oracle.com/big-data/what-is-big-data.html.

Free data sources

- **https://www.data.gov/**
- **https://www.census.gov/data.html**
- **https://data.gov.uk/**
- **http://data.europa.eu/euodp/en/data/**
- **https://developers.facebook.com/docs/graph-api**
- **https://www.healthdata.gov/**
- **http://content.digital.nhs.uk/home**
- **https://trends.google.com/trends/explore**
- **https://www.google.com/finance**
- **http://aws.amazon.com/datasets/**

Big Data Tools

- Apache Hadoop: Common & Modules

Apache Hadoop

- **The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.**

Extracted from: Apache Hadoop Docs & website: https://hadoop.apache.com/

Apache Hadoop

- **It's a framework for the distributed processing of large data sets across clusters of computers using simple programming models.**
- **Designed to scale up from single servers to thousands of machines, each offering local computation and storage.**

Extracted from: Apache Hadoop Docs & website: https://hadoop.apache.com/

Apache Hadoop

- **Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.**

Extracted from: Apache Hadoop Docs & website: https://hadoop.apache.com/

Hadoop Modules

- **Hadoop Common: The common utilities that support the other Hadoop modules.**
- **Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.**

Extracted from: Apache Hadoop Docs & website: https://hadoop.apache.com/

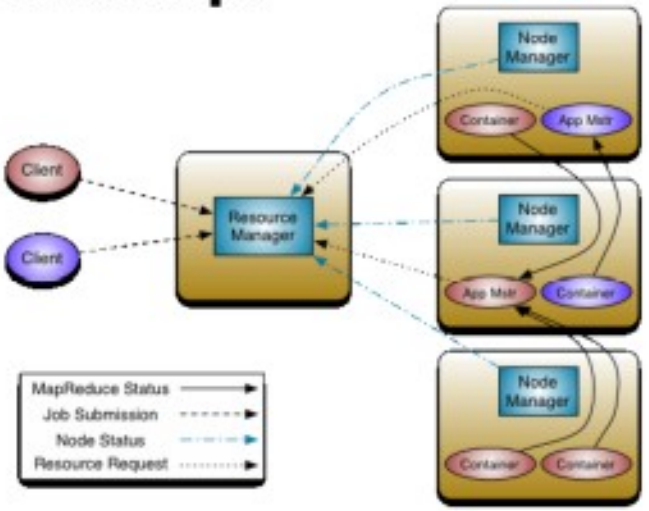
Hadoop Modules

- **Hadoop YARN: A framework for job scheduling and cluster resource management.**
- **Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.**
- **Hadoop Ozone: An object store for Hadoop.**

Extracted from: Apache Hadoop Docs & website: https://hadoop.apache.com/

Apache Hadoop YARN

- **Cluster management technology in second-generation Hadoop.**



Extracted from:
https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html

MapReduce

- **Software framework for processing massive amounts of unstructured data in parallel across a distributed cluster.**

Extracted from: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Apache Pig

- **Open source technology for parallel programming of MapReduce jobs on Hadoop clusters.**

Extracted from:https://pig.apache.org/

Apache Hadoop Tutorial

- **Hadoop installation on CentOS 8 Tutorial**
- **In this tutorial we'll install the Big Data framework Apache Hadoop on a previously installed CentOS 8 virtual machine.**
- **We'll use Docker containers for cluster creation.**
  - LINK:https://fortinux.com/linux-2-tutoriales/hadoop-installation-centos8/#more-1750

Questions?

Bibliography

SKIENA, Steven S. The Data Science Design Manual. (2017). Switzerland: Springer.

KUBAT, Miroslav. An Introduction to Machine Learning. Second Edition (2017). Switzerland: Springer.

TURKINGTON, Garry. Hadoop Beginner's Guide. (2013). UK: Packt Publishing.

O'REILLY RADAR TEAM. Planning for Big Data. (2012). USA: O'Reilly Media.

Copyright

Based on the following works:

SKIENA, Steven S. The Data Science Design Manual. (2017). Switzerland: Springer.

KUBAT, Miroslav. An Introduction to Machine Learning. Second Edition (2017). Switzerland: Springer.

TURKINGTON, Garry. Hadoop Beginner's Guide. (2013). UK: Packt Publishing.

O'REILLY RADAR TEAM. Planning for Big Data. (2012). USA: O'Reilly Media.

Copyright

FUNDAMENTALS OF BIG DATA

THANK YOU

FUNDAMENTALS OF BIG DATA

Data analysis & ML

Marcelo Horacio Fortino
MBA PM | PSM I | ITIL & ISO 20000
www.fortinux.com | @HoracioGRC