SAINT MARY'S UNIVERSITY SINCE 1802

One University. One World. Yours.

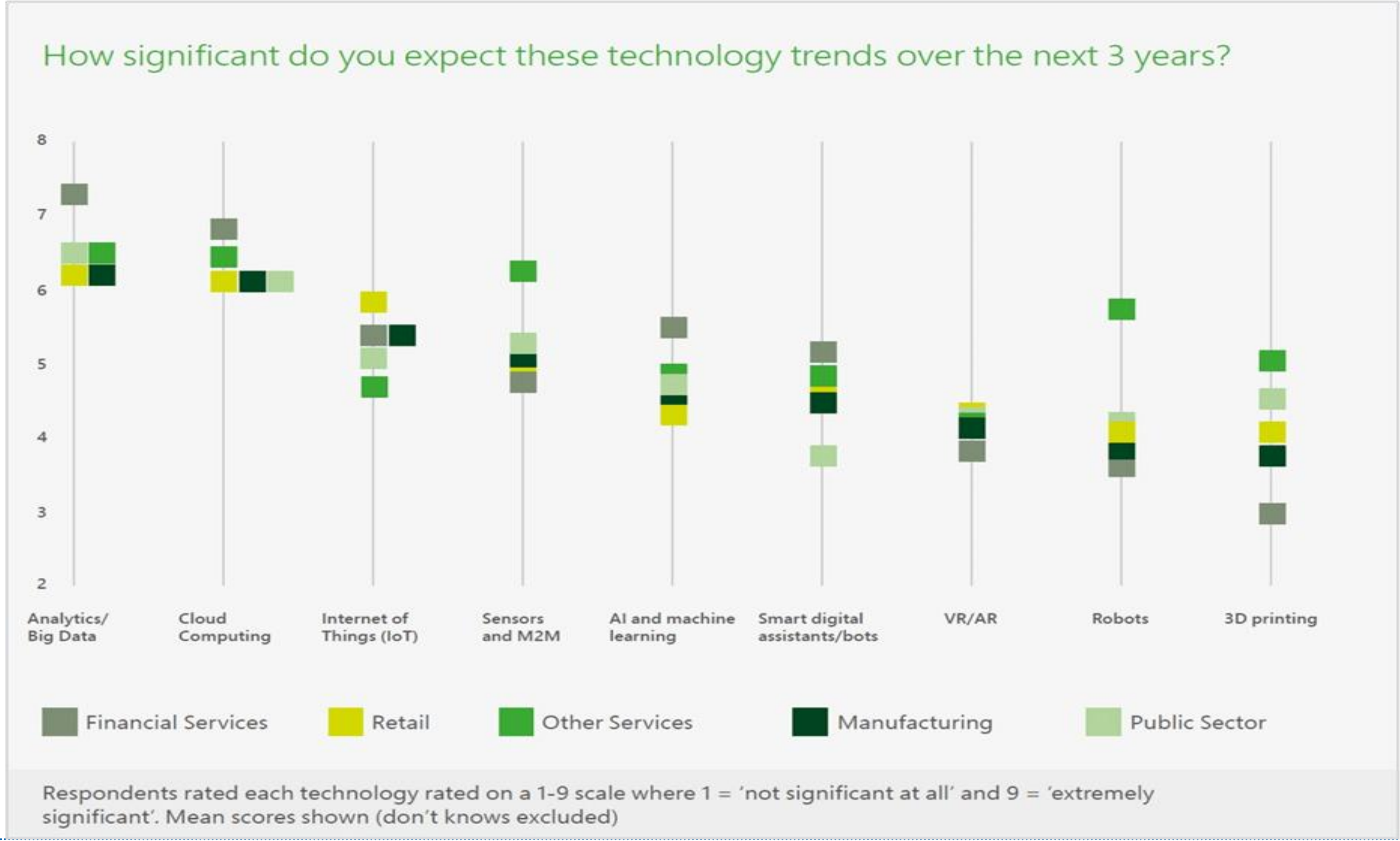
Big Data Workshop Presentation

- Cloud - introduction and comparison of vendors
- Hadoop - distributed databases and file systems
- Spark - distributed computing/data analytics

Gartner defines **cloud computing** as a style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using Internet technologies.

2018 Gartner IT Glossary > Cloud Computing

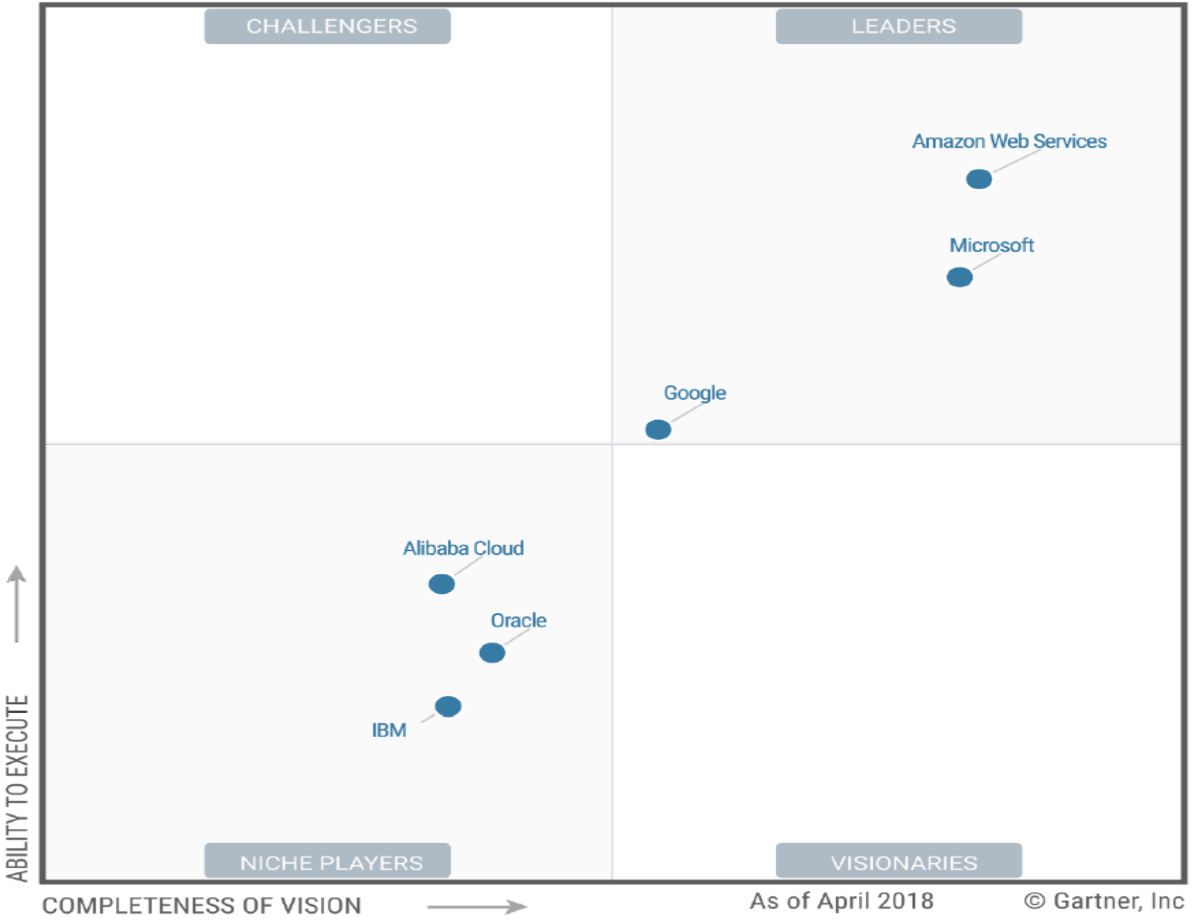
Analytics/Big Data & Cloud Computing Remain Top Trends



How significant do you expect these technology trends over the next 3 years?

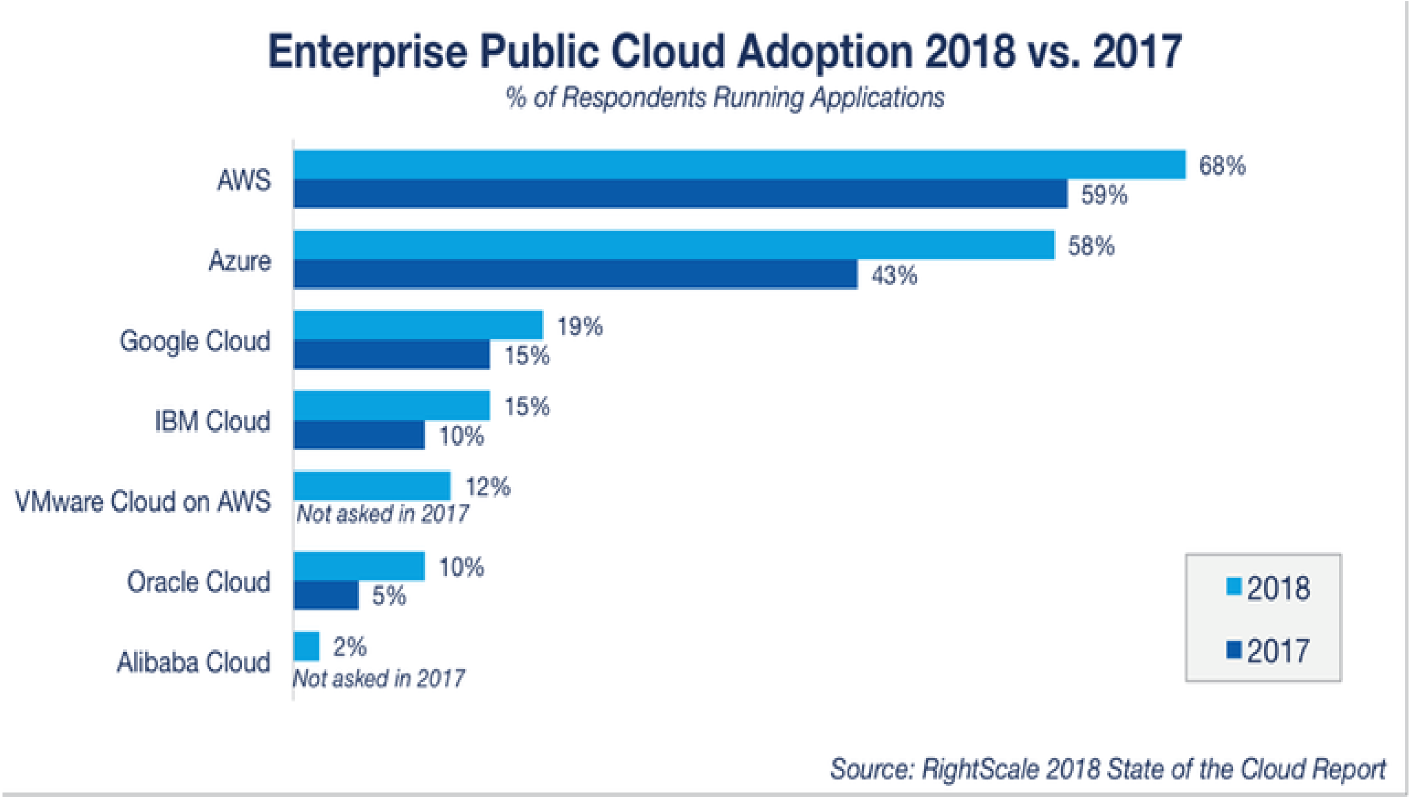
Respondents rated each technology rated on a 1-9 scale where 1 = 'not significant at all' and 9 = 'extremely significant'. Mean scores shown (don't knows excluded)

Gartner Magic Quadrant for Cloud Providers 2018



Source: Gartner (May 2018)

Stacking-Up-Cloud-Vendors 2018 vs 2017

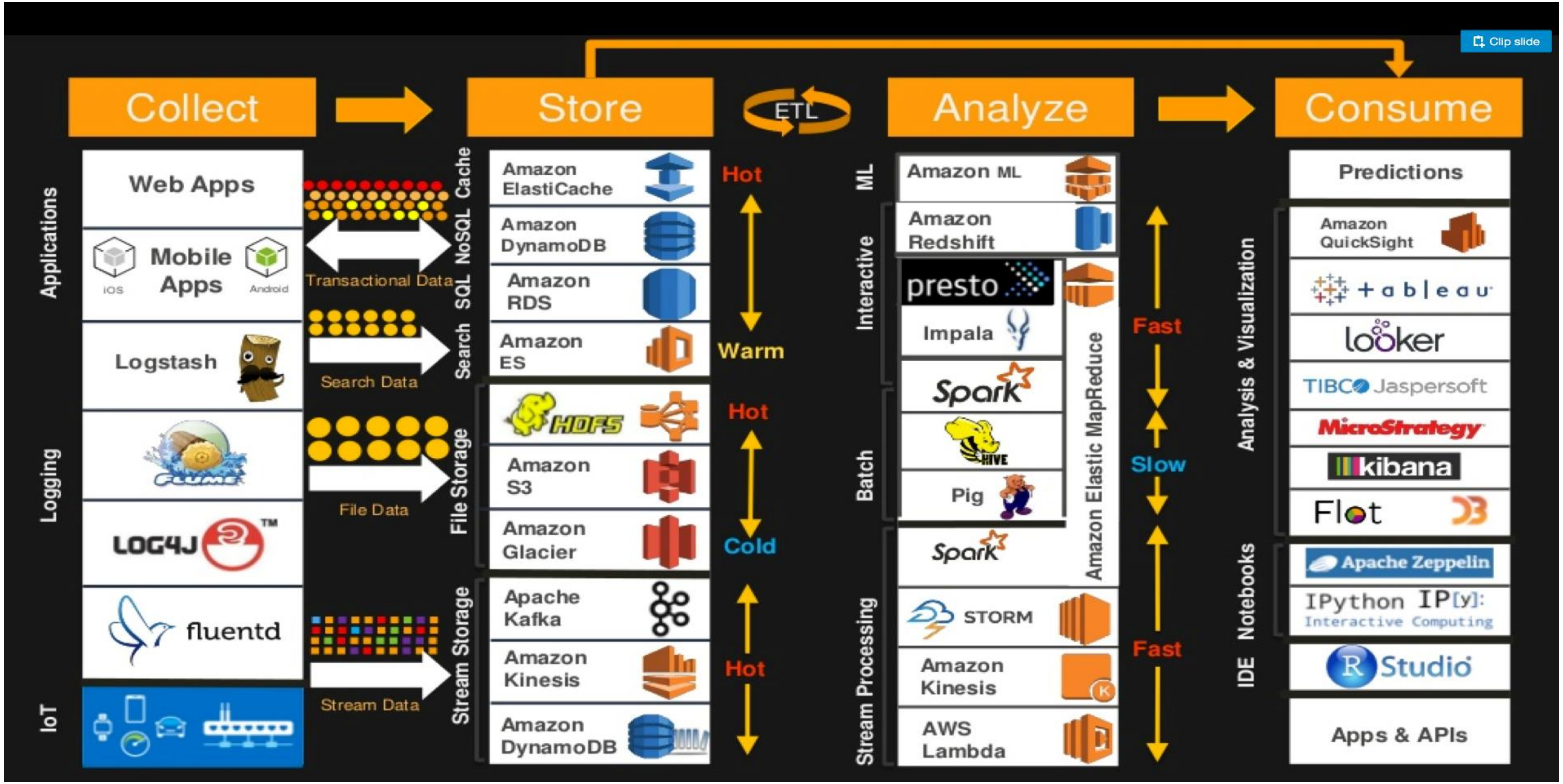


Enterprise Public Cloud Adoption 2018 vs. 2017

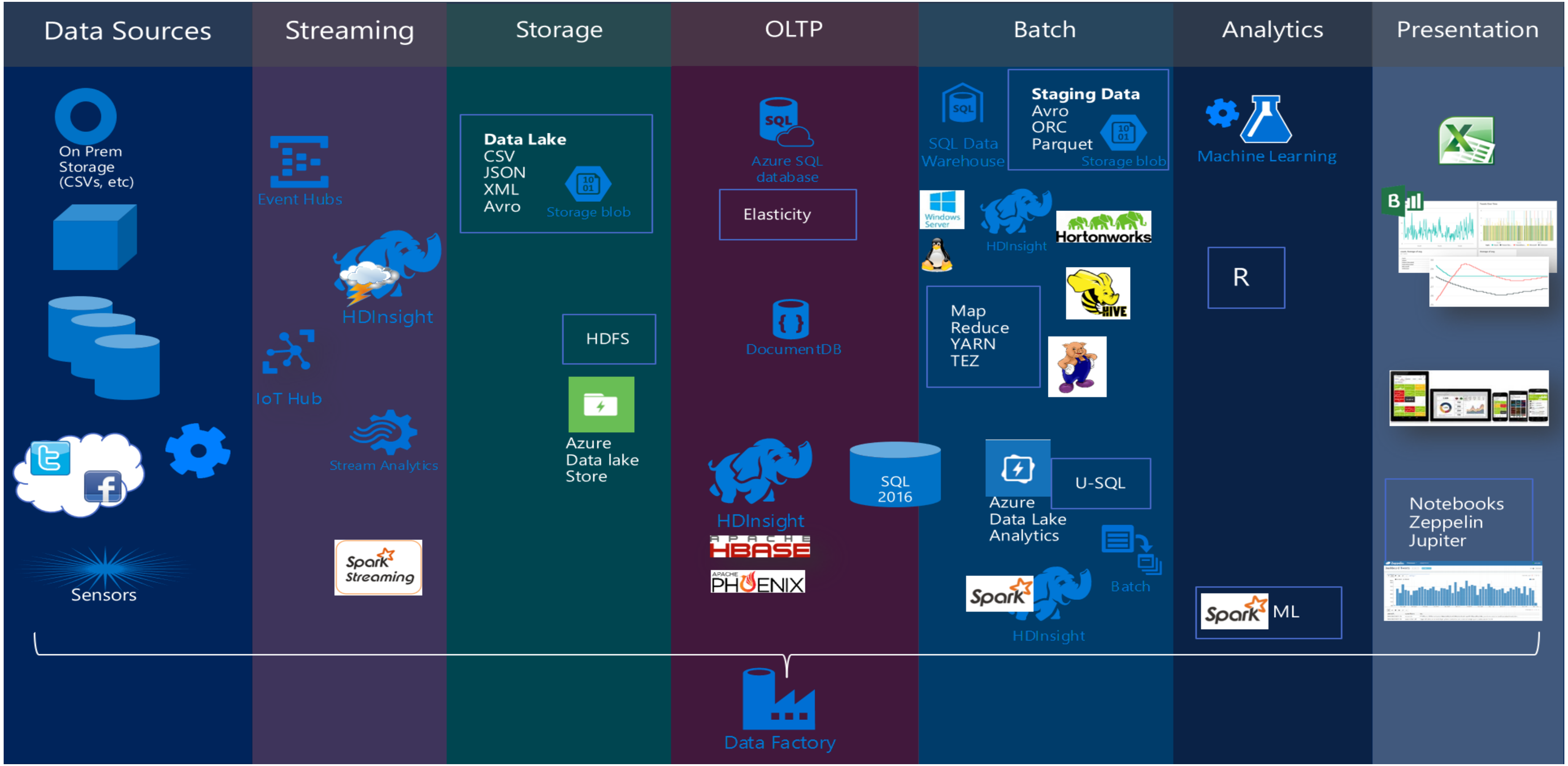
% of Respondents Running Applications

Source: RightScale 2018 State of the Cloud Report

AWS Big Data Reference Architecture



Azure Big Data Reference Architecture



Data Sources

On Prem Storage (CSVs, etc)

Sensors

Streaming

Event Hubs

HDInsight

IoT Hub

Stream Analytics

Spark Streaming

Storage

Data Lake
CSV
JSON
XML
Avro

Storage blob

HDFS

Azure Data lake Store

OLTP

Azure SQL database

Elasticity

DocumentDB

HDInsight

HBASE

PHOENIX

SQL 2016

Batch

SQL Data Warehouse

Staging Data
Avro
ORC
Parquet

Storage blob

Windows Server

HDInsight

Hortonworks

HIVE

Map
Reduce
YARN
TEZ

Azure Data Lake Analytics

U-SQL

Batch

Spark

HDInsight

Analytics

Machine Learning

R

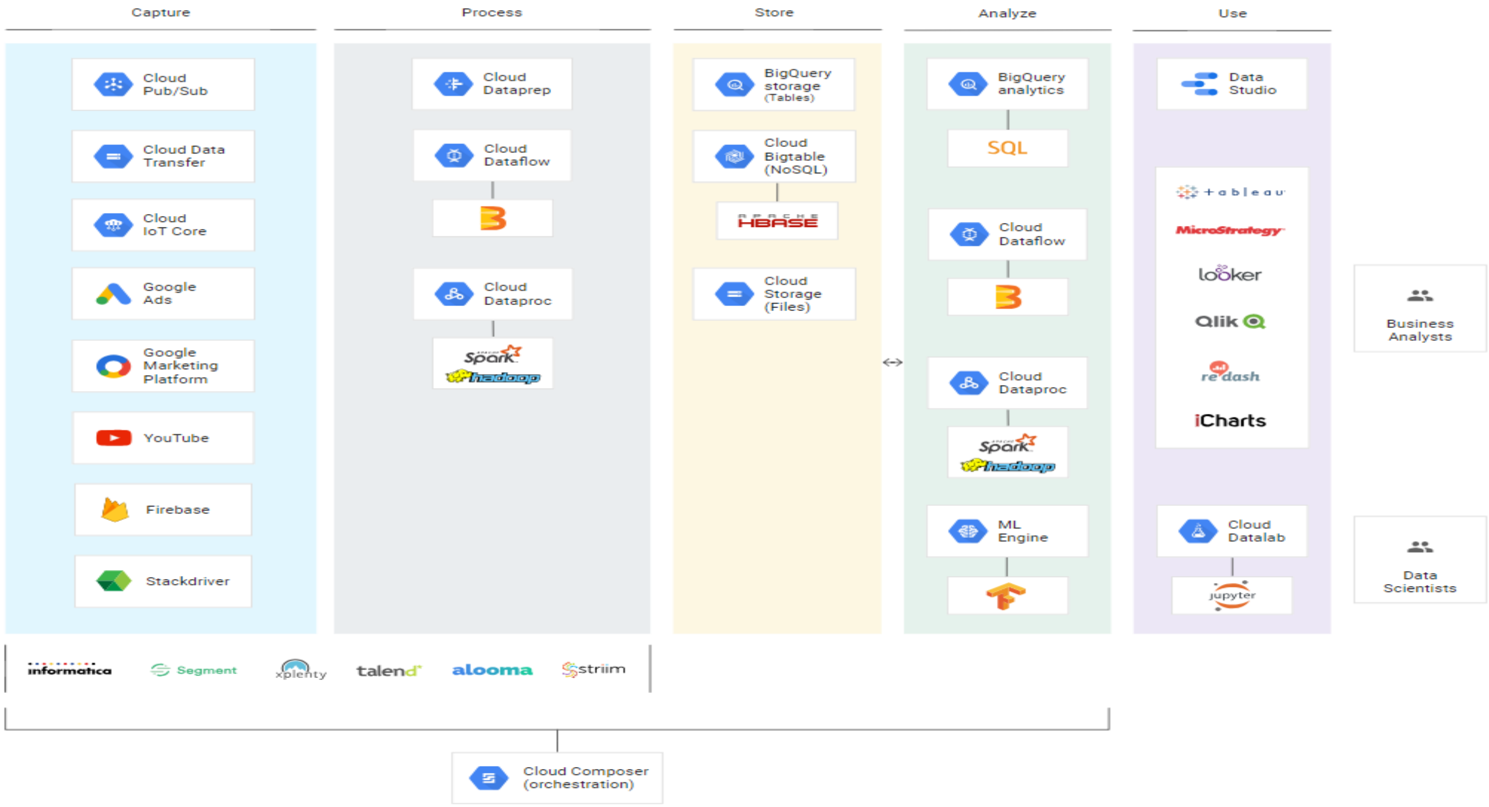
Spark ML

Presentation

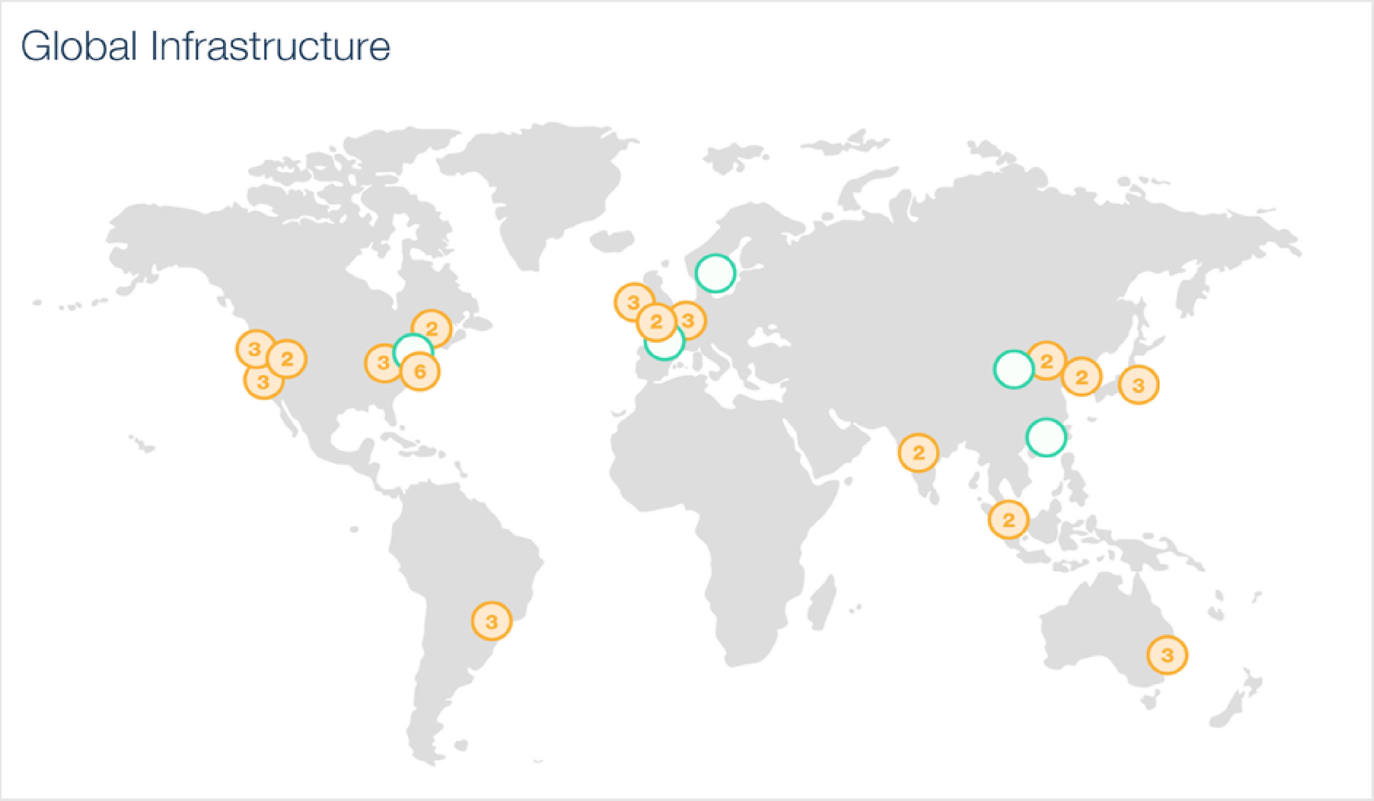
Notebooks
Zeppelin
Jupiter

Data Factory

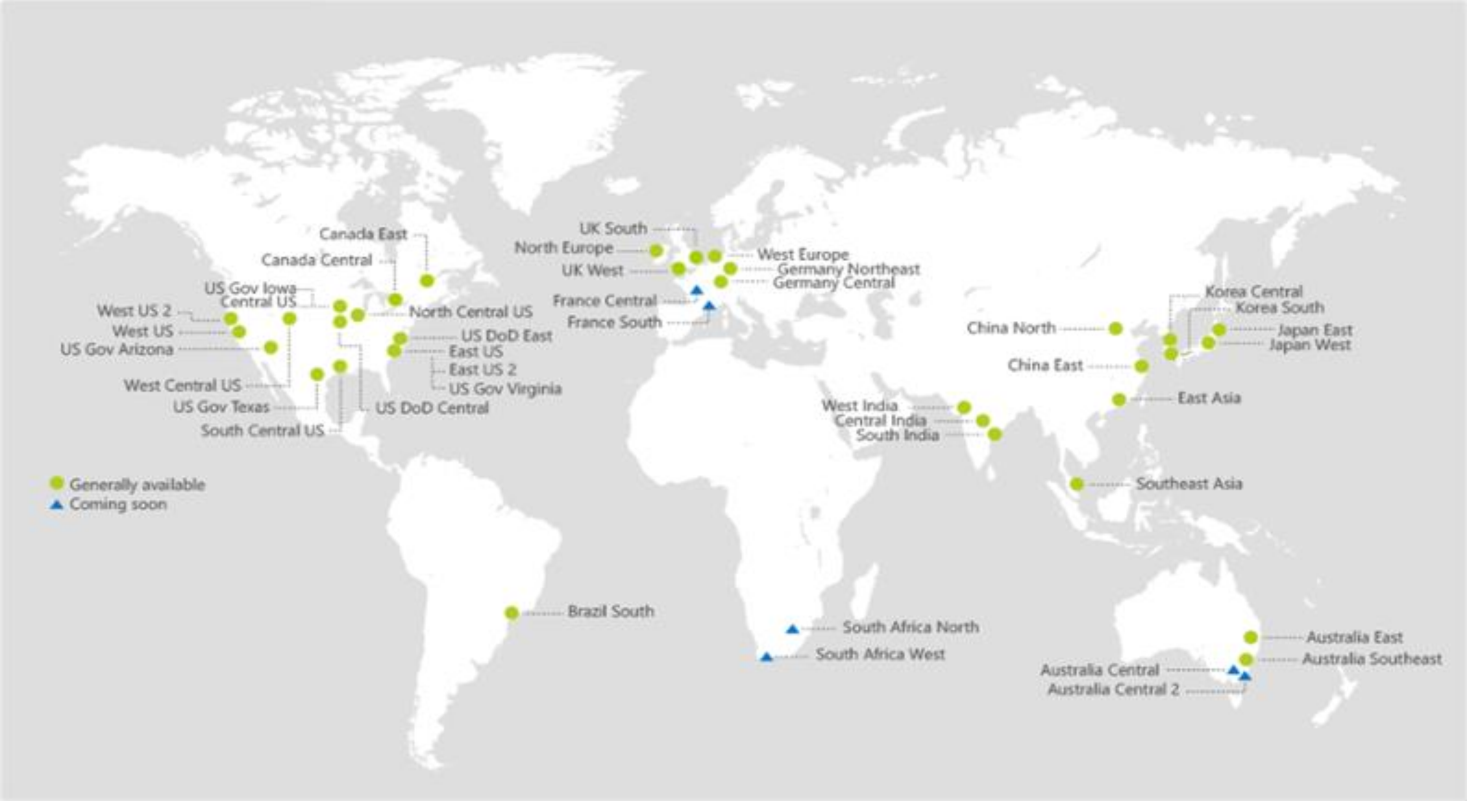
Google Big Data Reference Architecture (Serverless)



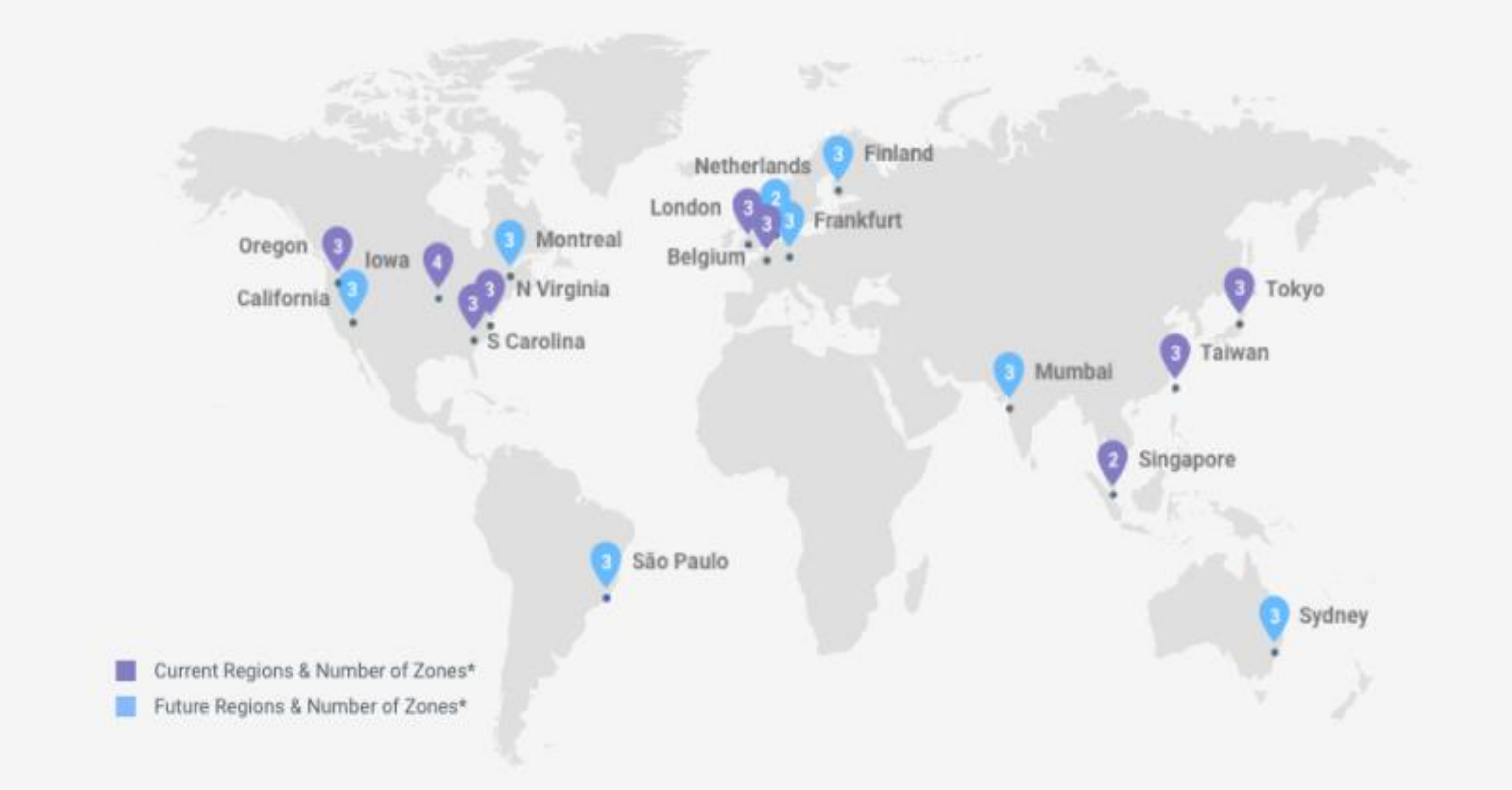
AWS Cloud Datacentres



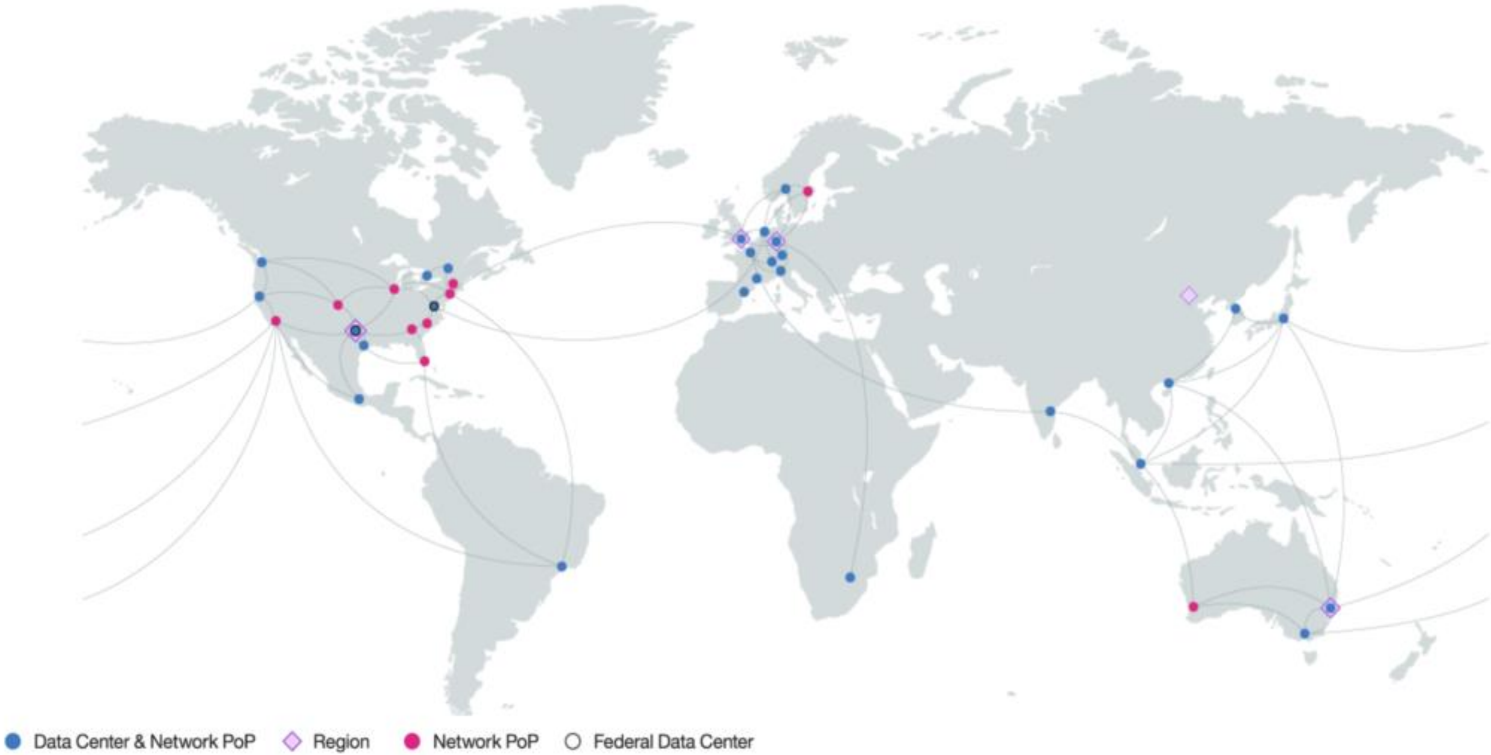
Azure Cloud Datacentres



Google Cloud Datacentres



IBM Cloud Datacentres



Cloud Summary

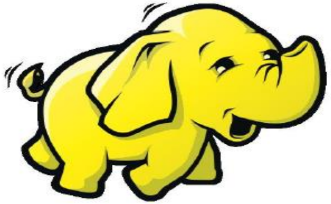
- The top two cloud vendors are AWS & Azure.
- AWS is currently the leader, Azure is gaining.
- Pricing models are based upon usage, able to pay as you go (monthly to per minute).

What is Big Data?

Gartner defines **Big Data** as 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.

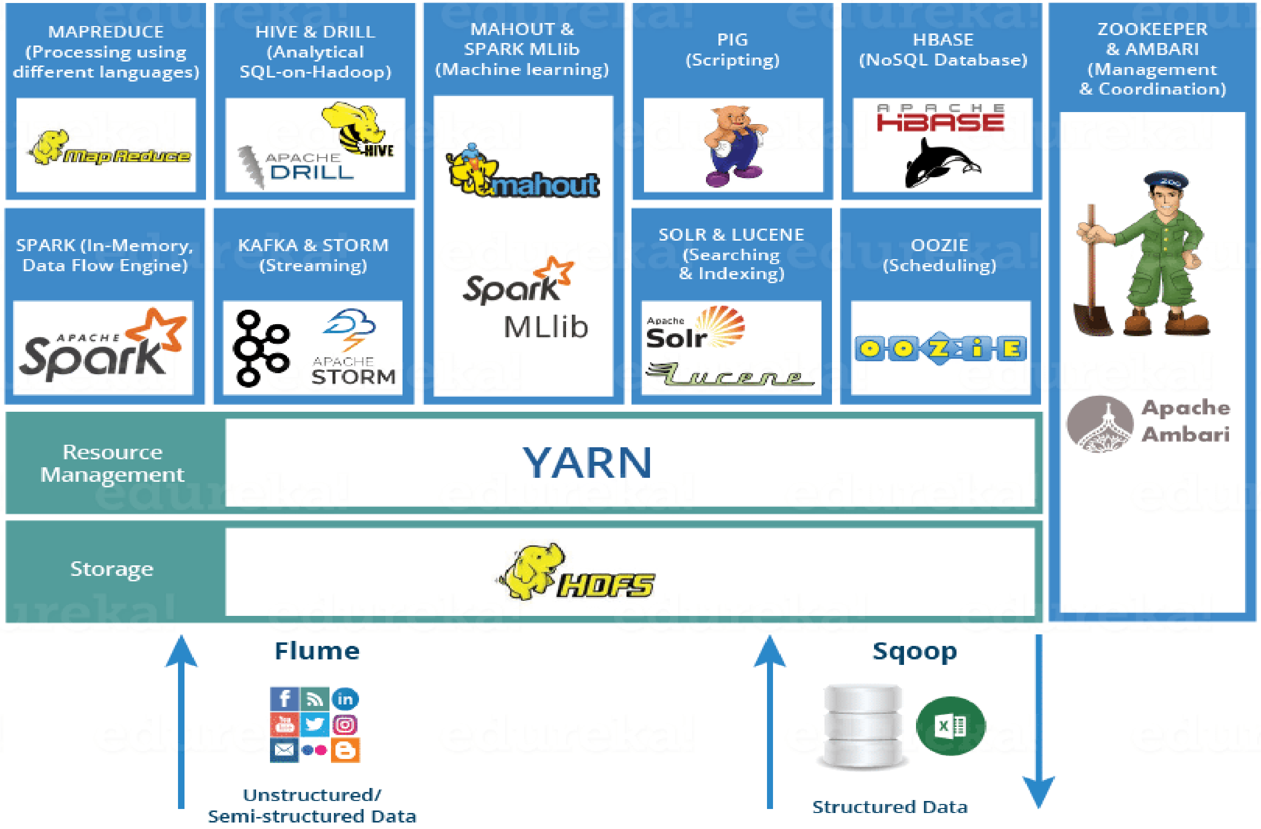
Gartner IT Glossary > Big Data

Hadoop



- Hadoop was inspired by the publication of Google's MapReduce , GoogleFS and BigTable . Hadoop was created by Doug Cutting and has been part of the Apache Software Foundation's projects since 2009.
- Today Hadoop is virtually synonymous with big data!

The Hadoop Ecosystem



MAPREDUCE (Processing using different languages)

HIVE & DRILL (Analytical SQL-on-Hadoop)

MAHOUT & SPARK MLlib (Machine learning)

PIG (Scripting)

HBASE (NoSQL Database)

ZOOKEEPER & AMBARI (Management & Coordination)

SPARK (In-Memory, Data Flow Engine)

KAFKA & STORM (Streaming)

SOLR & LUCENE (Searching & Indexing)

OOZIE (Scheduling)

Resource Management

YARN

Storage

Flume

Unstructured/ Semi-structured Data

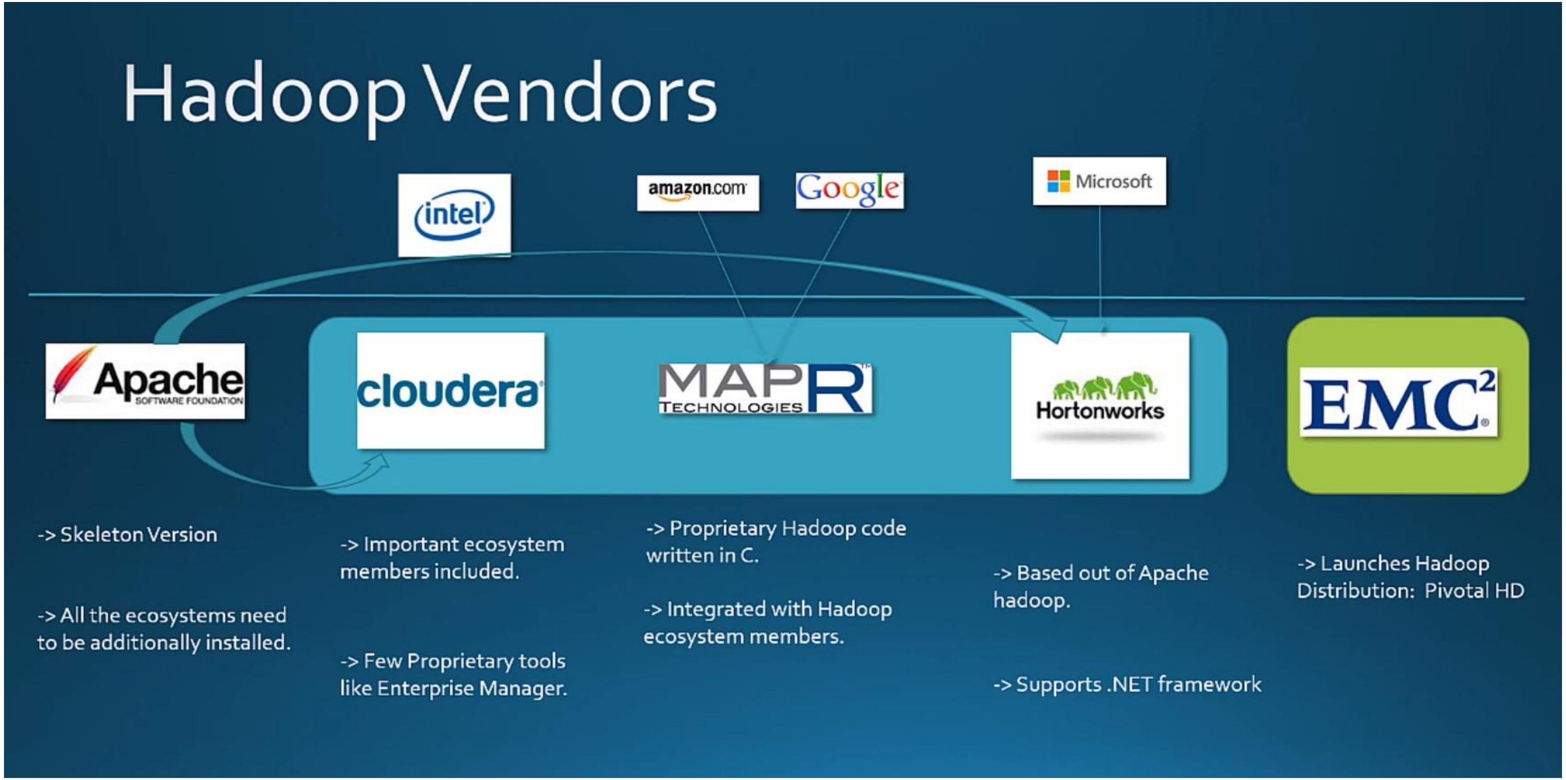
Sqoop

Structured Data

Deploying Hadoop

- Hadoop can be deployed in a traditional datacenter but also through cloud. The cloud enables organizations to deploy Hadoop without acquiring hardware or specific expertise.
- Azure HDInsight is a service that deploys Hadoop on Microsoft Azure. HDInsight uses Hortonworks Data Platform (HDP). HDInsight allows the programming of extensions in .NET (in addition to Java). HDInsight also supports the creation of Hadoop clusters using Ubuntu.
- Amazon Elastic Map Reduce (EMR) processes big data across a Hadoop cluster of virtual servers on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). The elastic in EMR's name refers to its dynamic resizing ability, which allows it to ramp up or reduce resource use depending on the demand at any given time.

Popular Hadoop Distributions



Hadoop Vendors

-> Skeleton Version

-> All the ecosystems need to be additionally installed.

-> Important ecosystem members included.

-> Few Proprietary tools like Enterprise Manager.

-> Proprietary Hadoop code written in C.

-> Integrated with Hadoop ecosystem members.

-> Based out of Apache hadoop.

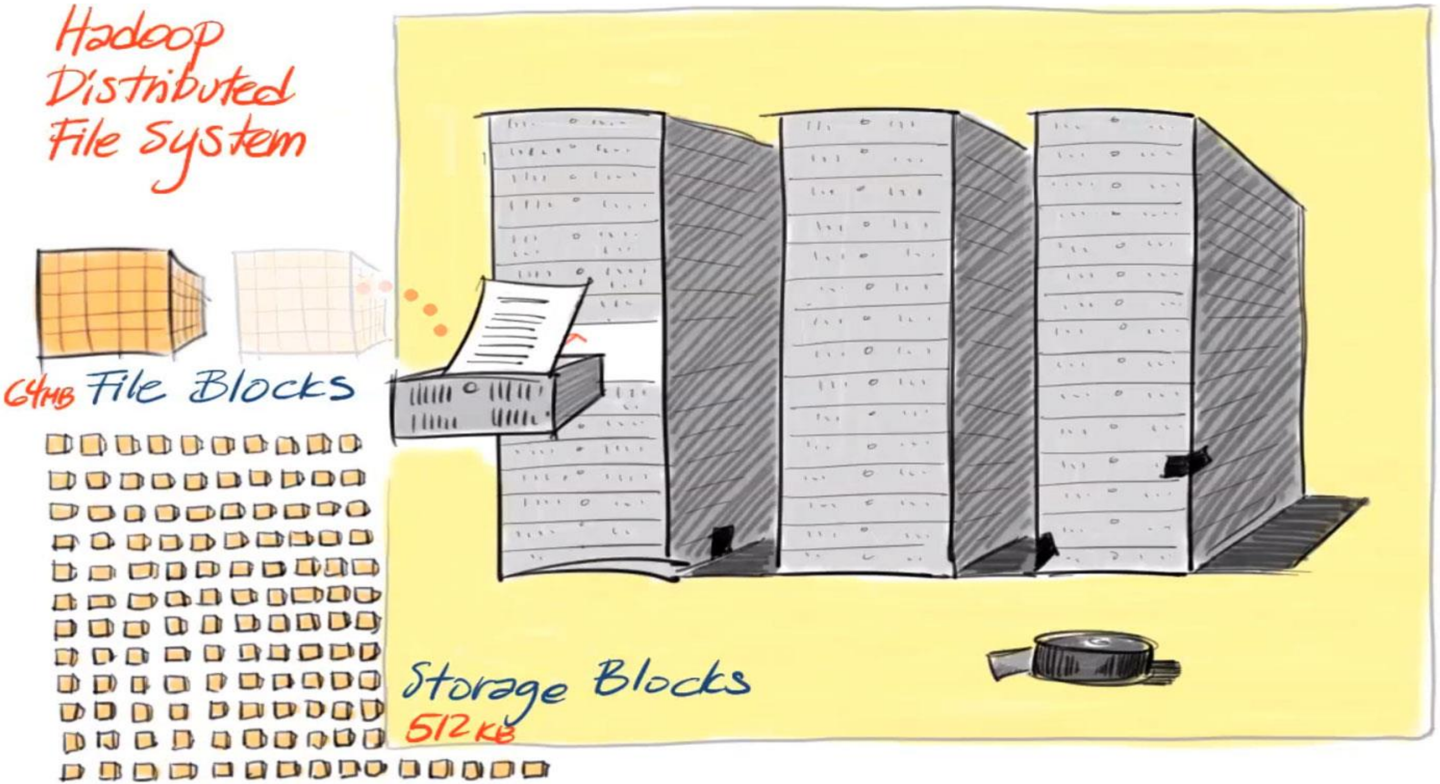
-> Supports .NET framework

-> Launches Hadoop Distribution: Pivotal HD

How the Hadoop Distributed File System (HDFS) Works

Hadoop has a file system that is much like the one on your desktop computer, but it allows us to distribute files across many machines. HDFS organizes information into a consistent set of file blocks and storage blocks for each node. In the Apache distribution, the file blocks are 64MB and the storage blocks are 512 KB. Most of the nodes are data nodes, and there are also three copies of the data. Name nodes exist to keep track of where all the file blocks reside.

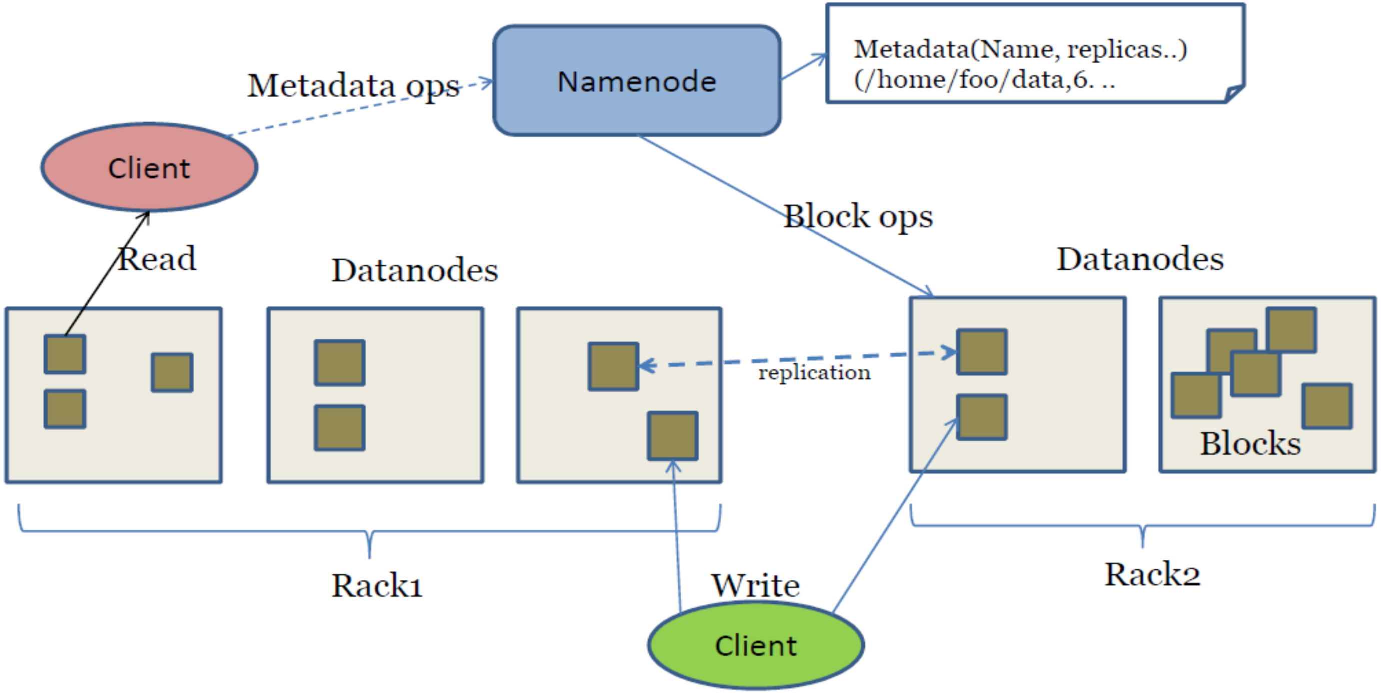
The Hadoop Distributed File System (HDFS)



Small Files in Hadoop

- Sometimes, you can get into trouble with small files on hdfs.
- A small file is one which is significantly smaller than the HDFS block size (default 64MB). If you’re storing small files, then you probably have lots of them (otherwise you wouldn’t turn to Hadoop), and the problem is that HDFS can’t handle lots of small files efficiently.
- To solve this problem, you should merge many of these small files into one and then process them.
- Solving the small files problem will shrink the number of map() functions executed and hence will improve the overall performance of a Hadoop job.

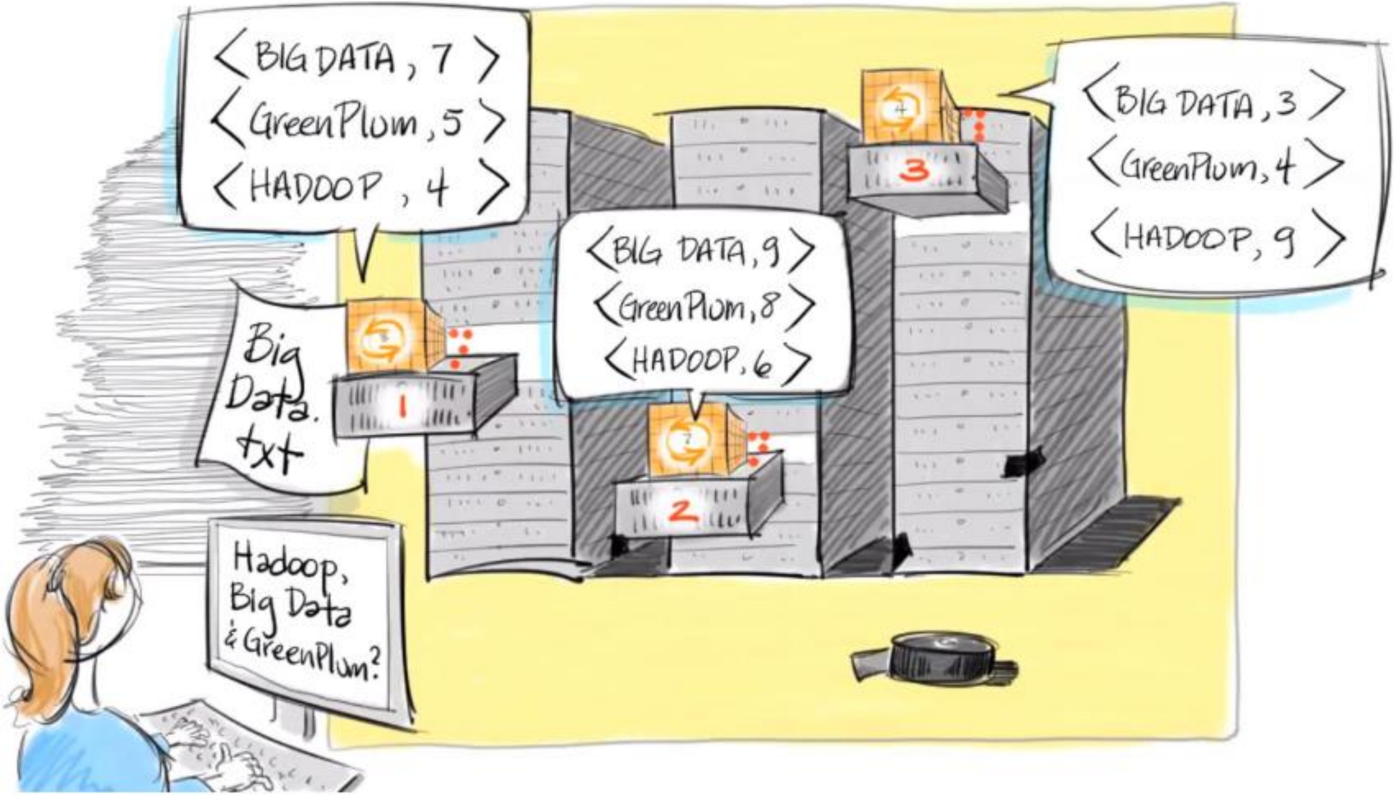
HDFS Architecture



How MapReduce Works

As the name suggests, there are two steps in the MapReduce process—map and reduce. Let’s say you start with a file containing all the blog entries about big data in the past 24 hours and want to count how many times the words “Hadoop”, “Big Data”, and “Greenplum” are mentioned. First, the file gets split up on HDFS. Then, all participating nodes go through the same map computation for their local dataset—they count the number of times these words show up. When the map step is complete, each node outputs a list of key-value pairs.

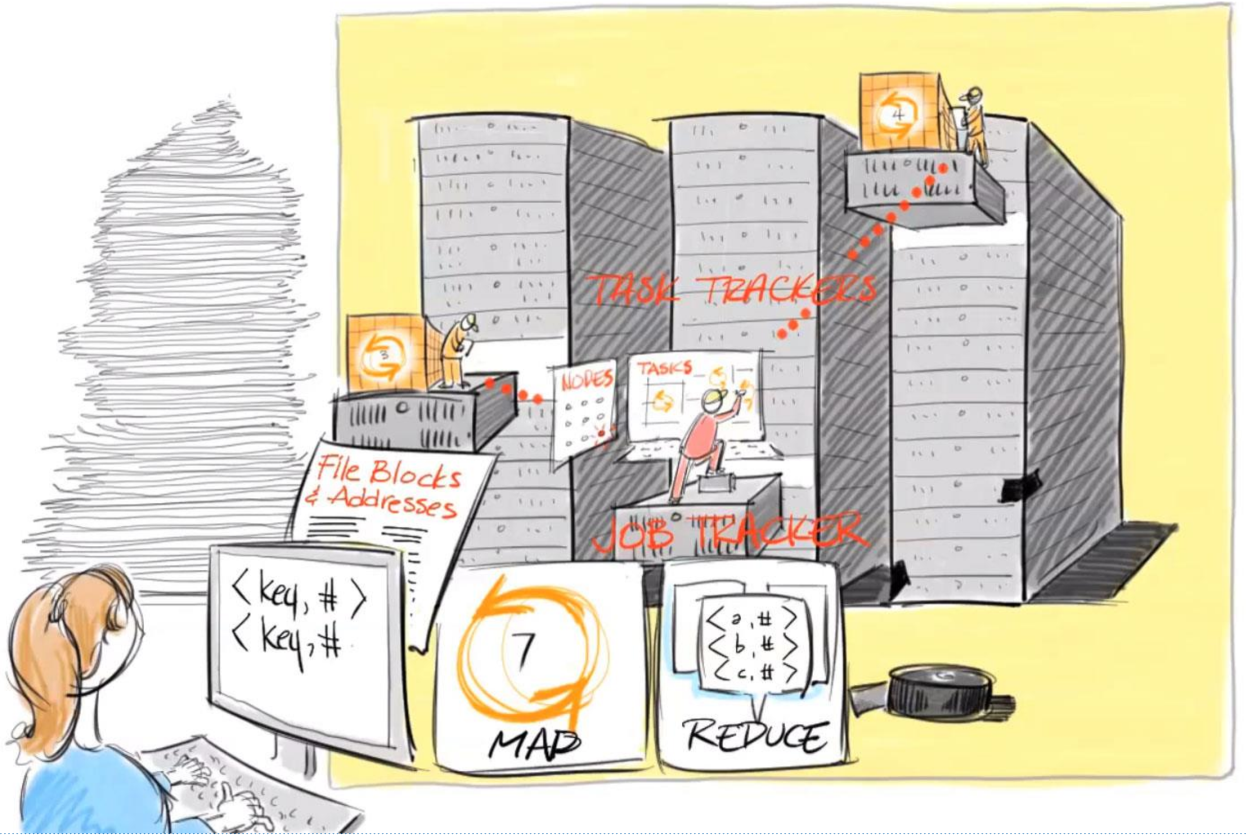
Hadoop MapReduce



Managing Hadoop Jobs

- If we had to split terabytes of data up by hand, copy the data to 1000 different computers manually, and kick each job off, the process would take forever and be prone to error. Fortunately, there is a set of components that automate all the steps.
- In Hadoop, the entire process is called a job, and a job tracker exists to divide the job into tasks and schedules tasks to run on the nodes. The job tracker keeps track of the participating nodes, monitors the processes, orchestrates data flow, and handles failures.
- Task trackers run tasks and report to the job tracker. With this layer of management automation, Hadoop can automatically distribute jobs on a large number of nodes in parallel and scale when more nodes are added.

Hadoop Job Task Trackers

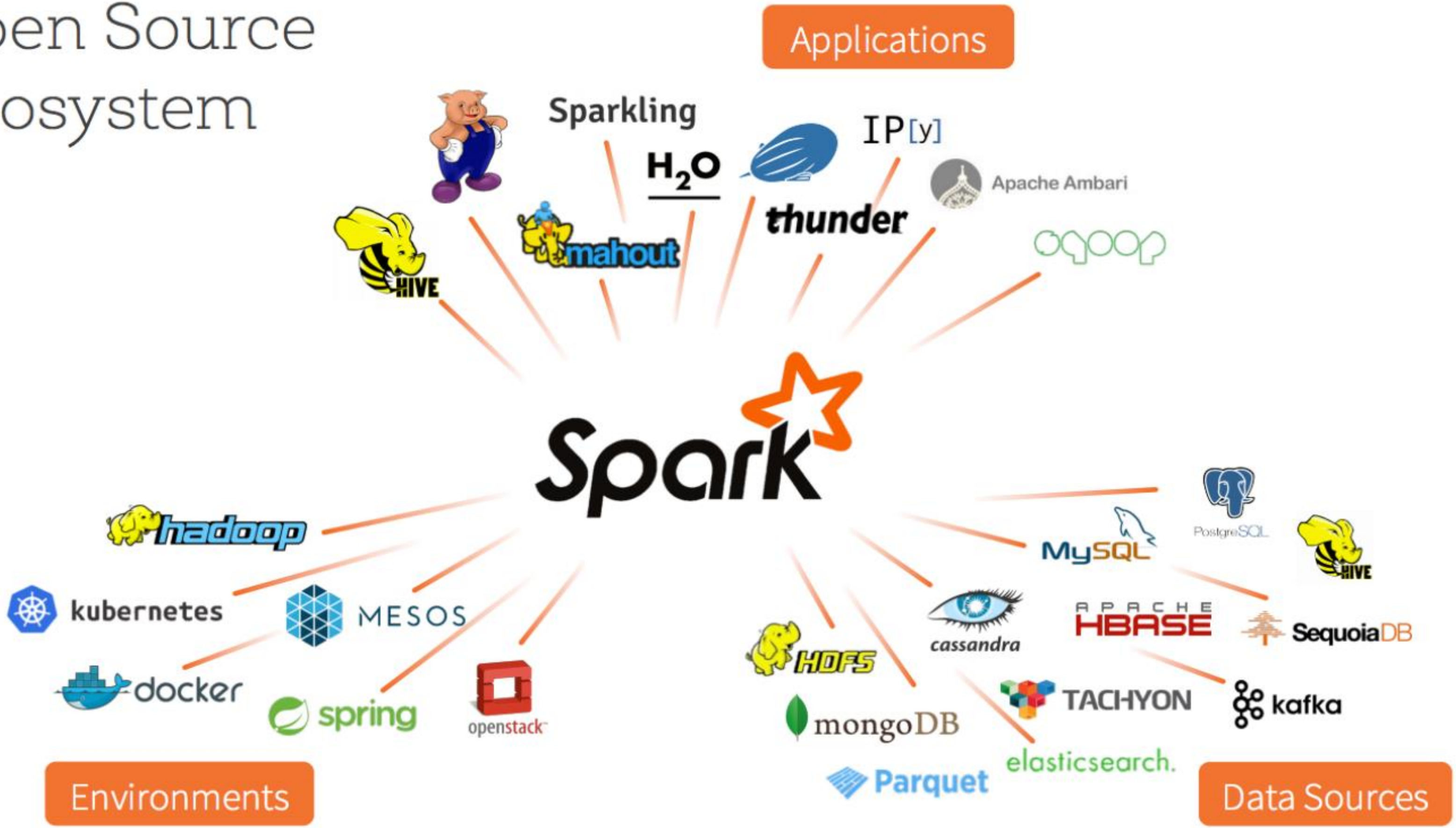


Distributed Computing/Data Analytics

“By 2020, smart, governed, Apache Hadoop/Spark-, search- and visual-based data discovery capabilities will converge into a single set of next-generation data discovery capabilities, as components of modern BI and analytics platforms.”

2018 Data and Analytics Programs Primer, Gartner

Open Source Ecosystem

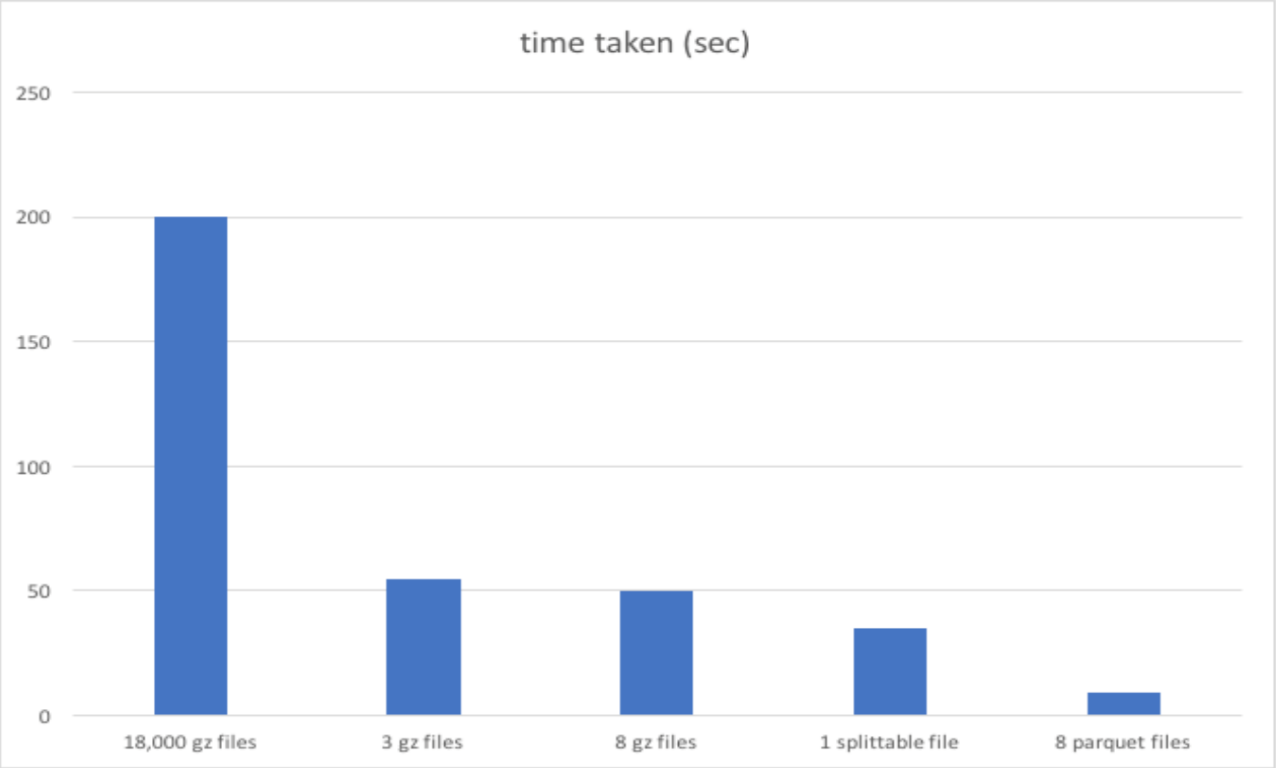


Apache Spark vs Hadoop

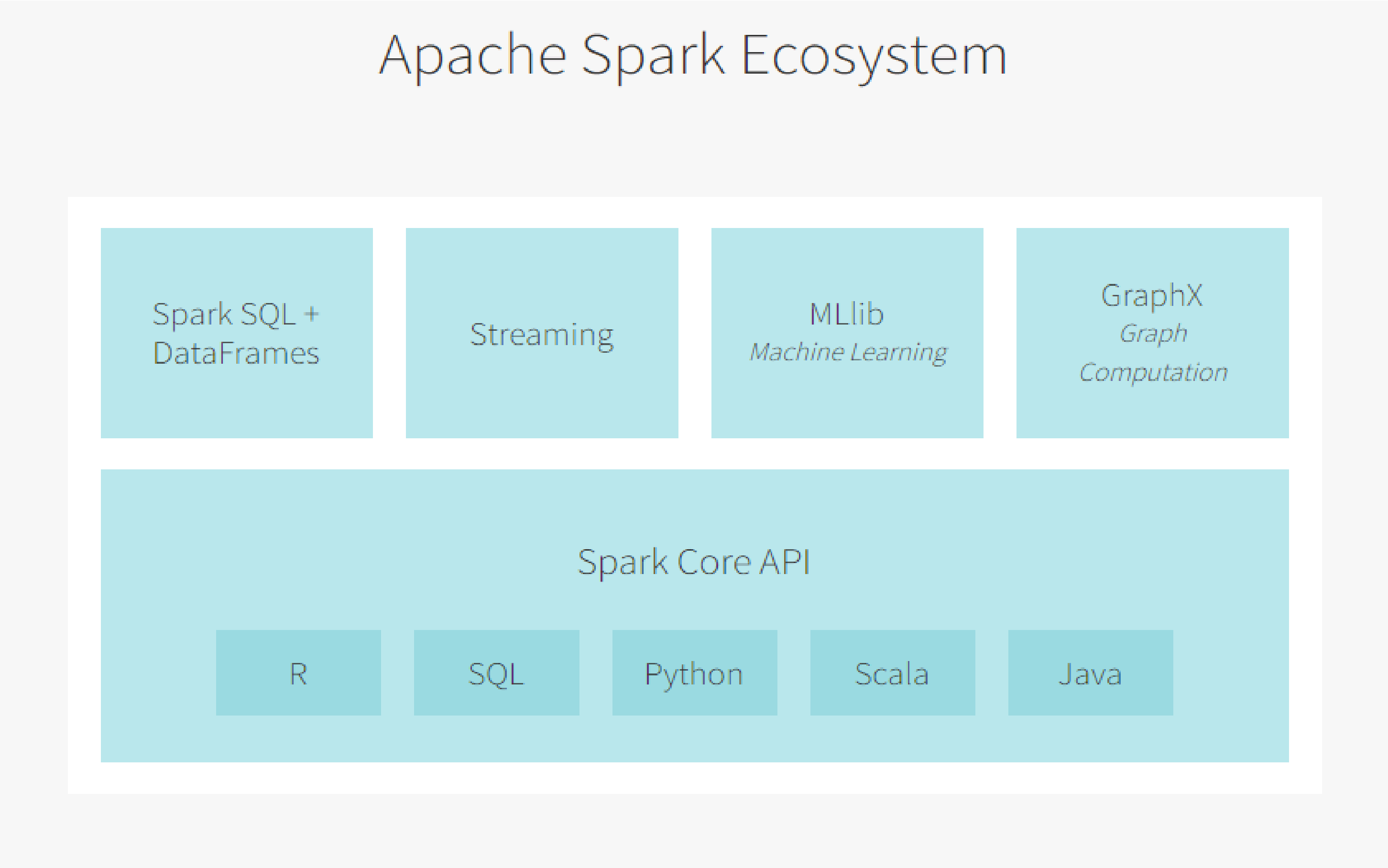
- Apache Spark is an open source, general-purpose distributed computing engine used for processing and analyzing large amounts of data. Like Hadoop MapReduce, Spark also works with the system to distribute data across the cluster and process the data in parallel.
- Spark is a cluster-computing framework, which means that it competes more with MapReduce than with the entire Hadoop ecosystem.
- Spark doesn't have its own distributed filesystem, but can use HDFS, S3, RDBMs, Elasticsearch, etc. Spark uses memory which makes it much faster but can use disk for processing, whereas MapReduce is strictly disk-based.
- As Spark typically stores data in memory if the dataset is very large and memory is limited it may be necessary to use disk or increase memory.

Apache Spark RDD

- Spark was developed in 2012 in response to the limitations of MapReduce (issue with small files, slow processing speed, support for batch processing only, not easy to use, no caching, etc.).
- Whereas Hadoop reads and writes files to HDFS, Spark processes data in memory using a concept known as an RDD, Resilient Distributed Dataset.
- The resilient distributed dataset (RDD), is a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way.



Using Scala with Spark to take advantage of Scala and Spark’s unique parallel job submission.



Apache Spark Ecosystem

Spark SQL + DataFrames

Streaming

MLlib *Machine Learning*

GraphX *Graph Computation*

Spark Core API

R

SQL

Python

Scala

Java

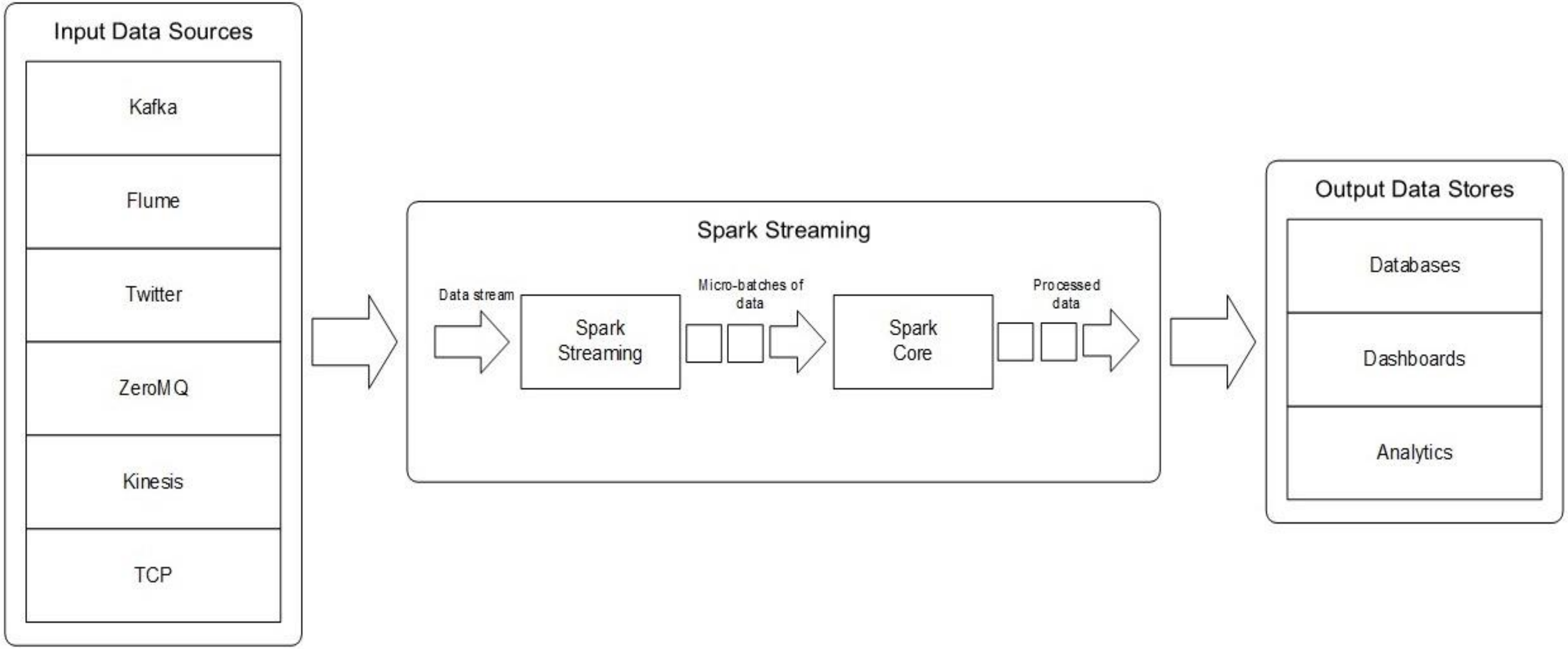
General Execution: Spark Core

Spark Core is the underlying general execution engine for the Spark platform that all other functionality is built on top of. It provides in-memory computing capabilities to deliver speed, a generalized execution model to support a wide variety of applications, and Java, Scala, and Python APIs for ease of development.

Streaming Analytics: Spark Streaming

Many applications need the ability to process and analyze not only batch data, but also streams of new data in real-time. Running on top of Spark, Spark Streaming enables powerful interactive and analytical applications across both streaming and historical data, while inheriting Spark’s ease of use and fault tolerance characteristics. It readily integrates with a wide variety of popular data sources, including HDFS, Flume, Kafka, and Twitter.

Spark Streaming Reference Architecture



Machine Learning: MLlib

Machine learning has quickly emerged as a critical piece in mining Big Data for actionable insights. Built on top of Spark, MLlib is a scalable machine learning library that delivers both high-quality algorithms (e.g., multiple iterations to increase accuracy) and blazing speed (up to 100x faster than MapReduce). The library is usable in Java, Scala, Python and R as part of Spark applications, so that you can include it in complete workflows.

Graph Computation: GraphX

GraphX is an extremely powerful graph computation engine built on top of Spark that enables users to interactively build, transform and reason about graph structured data at scale. It comes complete with a library of common algorithms.

Questions