













Inspire...Educate...Transform.

Applying ML to Big Data using Hadoop and Spark Ecosystem

Day 1: Big Data Overview, Hadoop Ecosystem, HDFS

Dr. Prasad M Deshpande

Slides: Dr. Manoj Duse

Applying Machine Learning to Big Data Using Hadoop and Spark

Foundations & Distributed Storage

Resource Management & Parallel Processing

More Spark and Spark ML

Streaming

• • • •

Agenda



- Big Data Overview
- Use Cases
- What is Hadoop, History/Evolution
- Hadoop Ecosystem
- HDFS



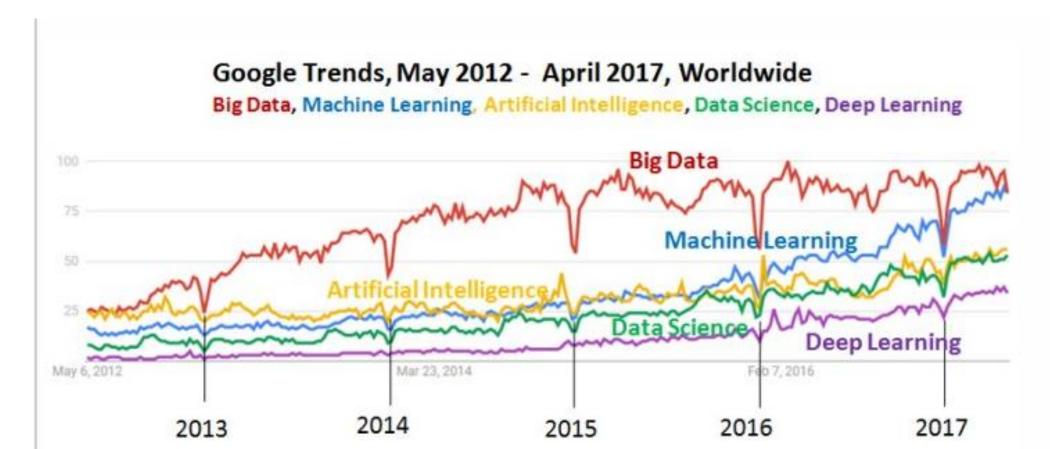


Fig. 3: Google Trends, May 2012 - April 2017, Worldwide

"Big Data" vs "Machine Learning" vs "Artificial Intelligence" vs "Data Science" vs "Deep Learning" search terms.

What is Big Data?



Big does not refer to size or volume alone!

So what else comes into play?

SUALIZATION

DECICION TO SUBMISTS ES PRINT





Volume

Sources:

Click stream

Logs

Social Media

IoT Sensors

Text Corpus [Blogs, Proposals, emails]





Velocity

- Speed at which data is generated
- Speed at which it has to be analysed for actionable insights

The V's of Big Data



Variety

- Structured
- Semi-structured
- Unstructured

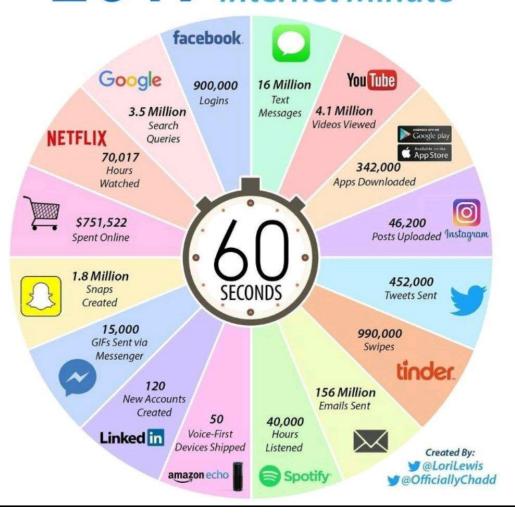
Veracity

- Trusted, Clean
- Untrusted, Uncleansed

Gartner, 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."

2017 This Is What Happens In An Internet Minute



2018 This Is What Happens In An Internet Minute



Orders of Magnitude

Name (Symbol)	Value	<u>Binary</u> <u>usage</u>
<u>kilobyte</u> (kB)	10 ³	2 ¹⁰
megabyte (MB)	10 ⁶	2 ²⁰
gigabyte (GB)	109	2 ³⁰
terabyte (TB)	10 ¹²	2 ⁴⁰
petabyte (PB)	10 ¹⁵	2 ⁵⁰
exabyte (EB)	10 ¹⁸	2 ⁶⁰
zettabyte (ZB)	10 ²¹	2 ⁷⁰
<u>yottabyte</u> (YB)	10 ²⁴	2 ⁸⁰

Let us understand the "scale" of data in today's world

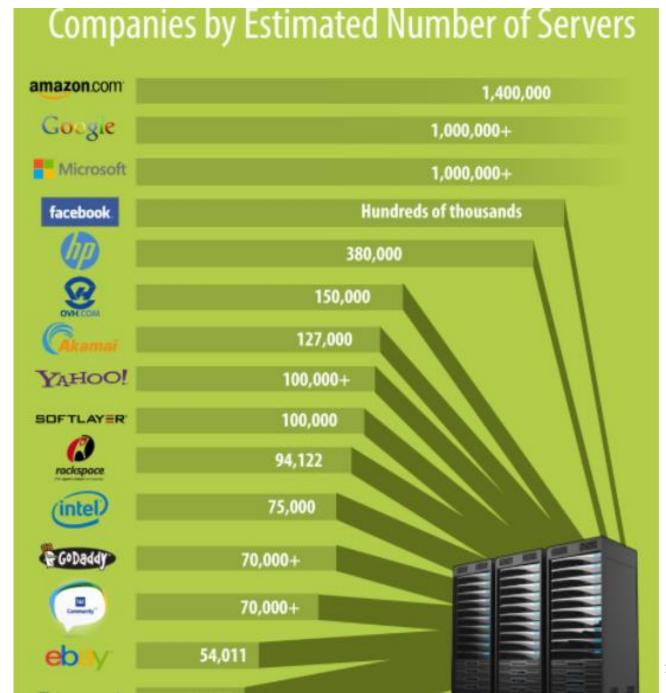
- Google is the largest 'big data' company in the world, processing 3.5 billion requests per day, storing 10 Exabytes (10 million terabytes) of data.
- Amazon hosts the most servers of any company, estimated at 1,400,000 servers with Google and Microsoft close behind.
- Amazon Web Services (AWS) are used by 60,000 companies and field more than 650,000 requests every second.
- Facebook collects 500 terabytes of data daily, including 2.5 billion pieces of content, 2.7 billion likes and 300 million photos.

• 90% of all the data in the world was produced in the last 2 years.

• It is estimated that 40 zettabytes (40,000 Exabytes) of data will be created by 2020. 3.2 zettabytes [in 2014]

 The total amount of data being captured and stored by industry doubles every 1.2 years

How big is Big?



○ 10 April 2014

◆ DATA · INFRA · PRODUCTION ENGINEERING

Scaling the Facebook data warehouse to 300 PB



At Facebook, we have unique storage scalability challenges when it comes to our data warehouse. Our warehouse stores upwards of 300 PB of Hive data, with an incoming daily rate of about 600 TB. In the last year, the warehouse has seen a 3x growth in the amount of data stored. Given this growth trajectory, storage efficiency is and will continue to be a focus for our warehouse infrastructure.

Is all this data really useful?

• The more data you "do" data science with, the finer (and better) your insights will be.

What are big data use cases?

Big Data at CERN

- https://www.youtube.com/watch?v=j-0cUmUyb-Y
- Physicists at CERN have been pondering how to store and share their ever more massive data for decades - stimulating globalization of the internet along the way, whilst 'solving' their big data problem.
- Tim Smith plots CERN's involvement with big data from fifty years ago to today.
- Lesson by Tim Smith, animation by TED-Ed.

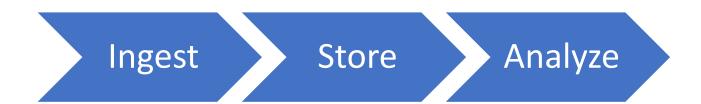
Some use cases

- Credit card fraud detection: Need transition graph analysis for discovering patterns in sequence of transactions
 - Example: detect Outlier Transactions with respect to space/time constraints
- Business optimization using Location based services
 - Real time connection of customers to merchants (e.g., providing shopping incentives, such as discounts)
- Customer analytics segmentation, churn

One of the Google data centres powering the cloud

https://www.youtube.com/watch?v=zDAYZU4A3w0

Big data processing



What are the design goals of a big data platform?

- Fast processing
 - Data ought to be in primary storage, or even better, RAM
- Scalable
 - Should be able to handle growing data volumes
- Reliable
 - Should be able to handle failures gracefully
- Ease of programming
 - Right level of abstractions to help build applications
- Low cost

➤ Need a whole ecosystem

Enter the world of "Hadoop"

Hadoop is Apache Project

What is Apache ...ASF?

WHAT IS THE APACHE SOFTWARE FOUNDATION?

The Apache Software Foundation (ASF) is a non-profit corporation [incorporated in 1999]

The ASF is a natural outgrowth of The Apache Group, a group of individuals that was initially formed in 1995 to develop the Apache HTTP Server.

The all-volunteer ASF develops, stewards, and incubates more than 350 Open Source projects and initiatives that cover a wide range of technologies.

WHY WAS THE APACHE SOFTWARE FOUNDATION CREATED?

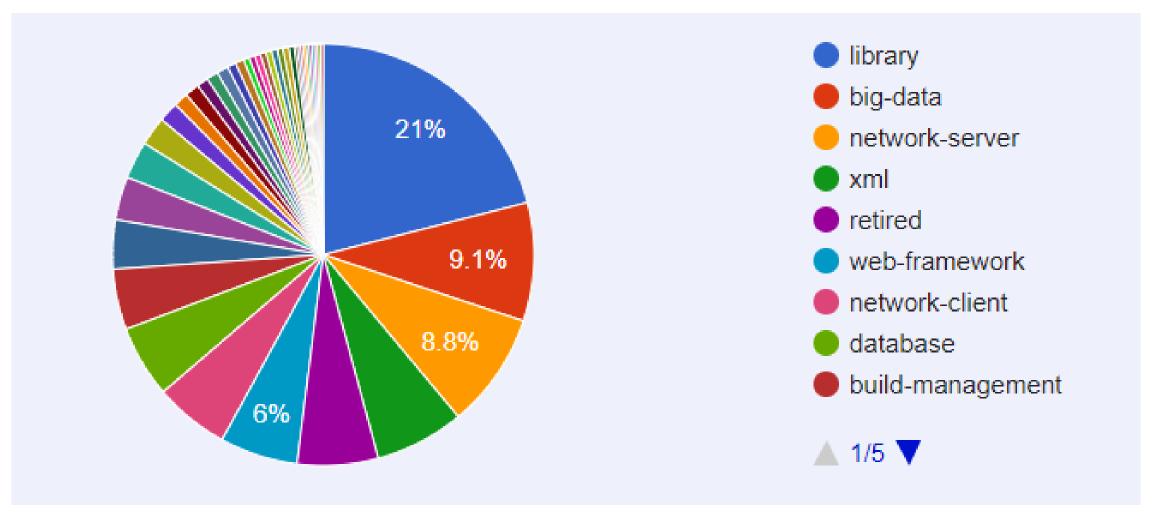
- 1.provide a foundation for open, collaborative software development projects
- 2.create an independent legal entity to which companies and individuals can donate resources and be assured that those resources will be used for the public benefit;
- 3.provide a means for individual volunteers to be sheltered from legal suits directed at the ASF projects;
- 4.protect the 'Apache' brand, as applied to its software products, from being abused

WHY WAS THE NAME 'APACHE' CHOSEN?

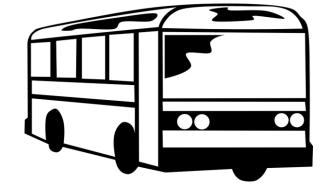
The name 'Apache' was chosen from respect for the various Native American nations collectively referred to as <u>Apache</u>, well-known for their superior skills in warfare strategy and their inexhaustible endurance.

It also makes a cute pun on "a patchy web server" -- a server made from a series of patches

Big Data Share in Apache Projects



How to scale?



40 passengers

Requirements

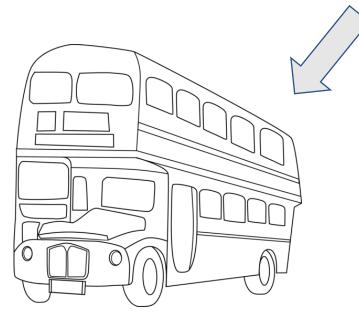
Speed

Scalability

Reliability

Ease of programming

Low cost





Horizontal Scaling

80 passengers

Vertical

Scaling

80 passengers

Ways to Scale

• To scale horizontally (or scale out) means to add more nodes to a system, such as adding a new computer to a distributed software application.

 To scale vertically (or scale up) means to add resources to a single node in a system, typically involving the addition of CPUs or memory to a single computer.

What are the advantages and disadvantages?

Challenges with Distributed Systems

- Complexity of
 - setup and administration
 - Programming parallel programs are very hard to write
- More points of failure
 - Node failure
 - Disk failure
 - Network failure

What is Hadoop?

 Hadoop is an open source framework, from the Apache foundation, capable of processing large amounts of heterogeneous data sets in a distributed fashion across clusters of commodity computers and hardware using a simplified programming model.

 The Hadoop framework is based closely on the following principle:

In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers. **~Grace Hopper**

Hadoop

HDFS - Reliable Shared Storage

+

MapReduce - Distributed Computation

Parallel processing

Hide the complexity of distributed computation

History of Hadoop

Hadoop was created by Doug Cutting and Mike Cafarella.

 Originated from an open source web search engine called "Apache Nutch", which is part of another Apache project called "Apache Lucene"

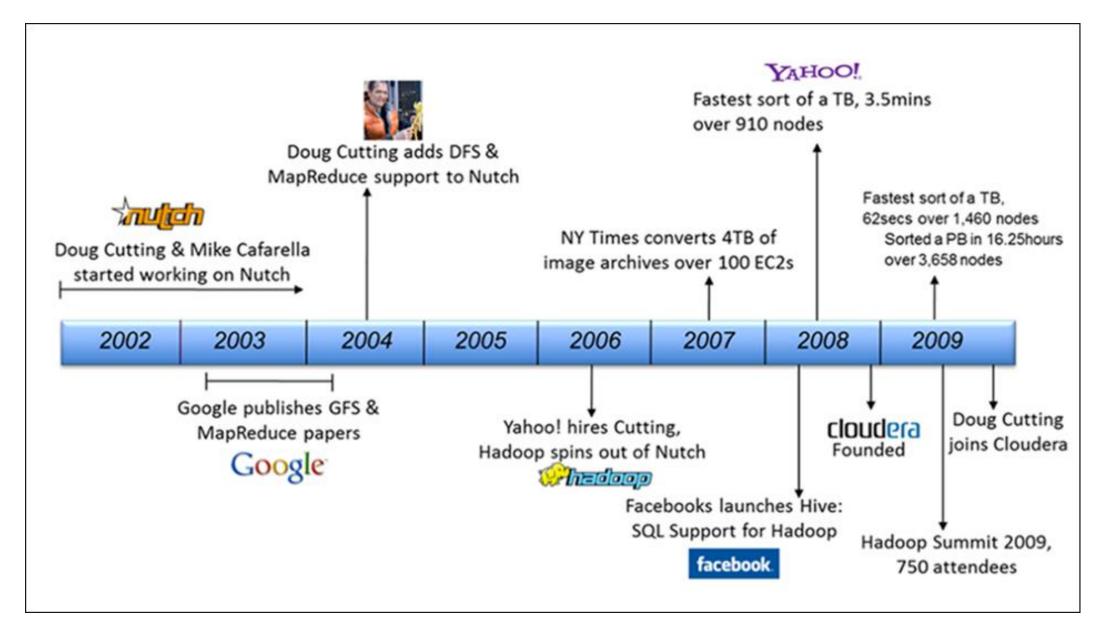
What does HADOOP stand for ?

 According to Hadoop's creator Doug Cutting, the name came about as follows:

•

"The name my kid gave a stuffed yellow elephant. Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria."

Hadoop Timeline



2011: HW And MapR

Doug Cutting Basics of Hadoop Video

https://www.youtube.com/watch?v=0GOxDBR6VAU





Some basic terms:

- Latency
- Throughput
- Bandwidth
- ✓ Scale Up (Vertical scaling), Scale Out (Horizontal Scaling)
- Serialization/Deserialization
- Rack, Switch

 Bandwidth: How wide the pipe is – theoretical maximum rate of flow/processing

 Latency: How long does it take to travel from one end of pipe to the other

Throughput: Actual amount flowing through per unit time

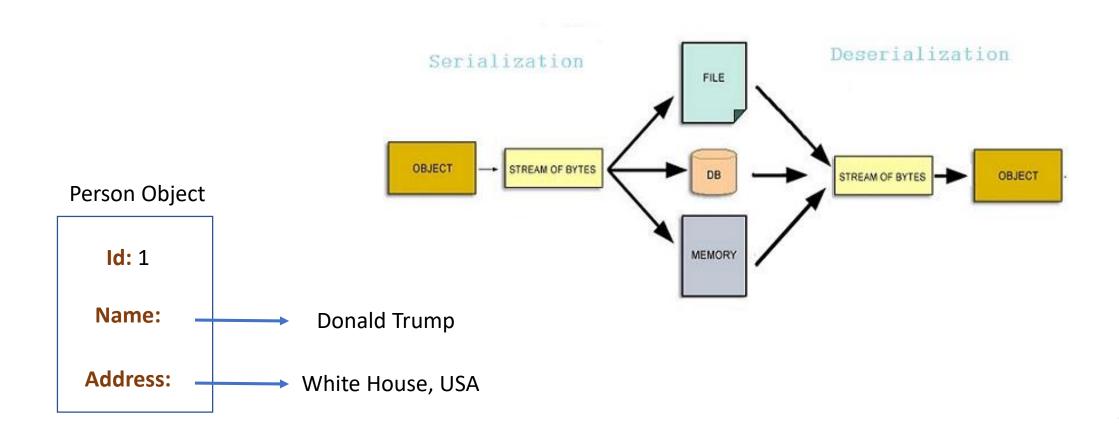
SerDe



• **Ser**ialization and **De**serialization

• What is serialization?

How will you store an object on disk?



Serialization

 Serialization is the process of converting the state information of an object instance into a binary or textual form to persist into storage medium or transported over a network.

Worded differently...

- Serialization is the process of converting an object into a stream of bytes in order to store the object or transmit it to memory, a database, or a file. [save]
- Its main purpose is to save the state of an object in order to be able to recreate it when needed.
- The reverse process is called deserialization. [restore]

Rack

The rack contains multiple mounting slots called bays, each designed to hold a hardware unit secured in place with screws.

A single rack can contain multiple servers stacked one above the other, consolidating network resources and minimizing the required floor space.

The rack server configuration also simplifies cabling among network components.

Cooling systems become critical aspects



Switch

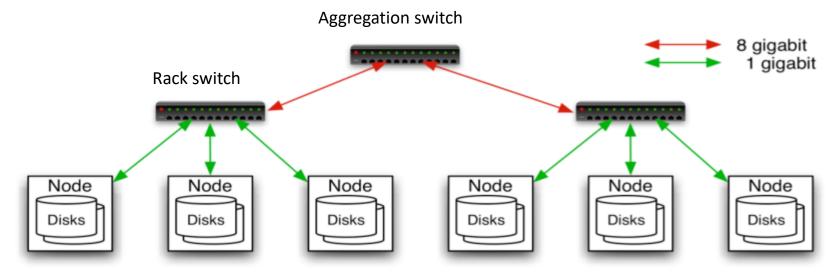
• A switch, in the context of networking is a high-speed device that receives incoming data packets and redirects them to their destination on a local area network (LAN).

• Essentially, switches are the traffic cops of a simple local area network

 Switch is limited to node-to-node communication on the same network.

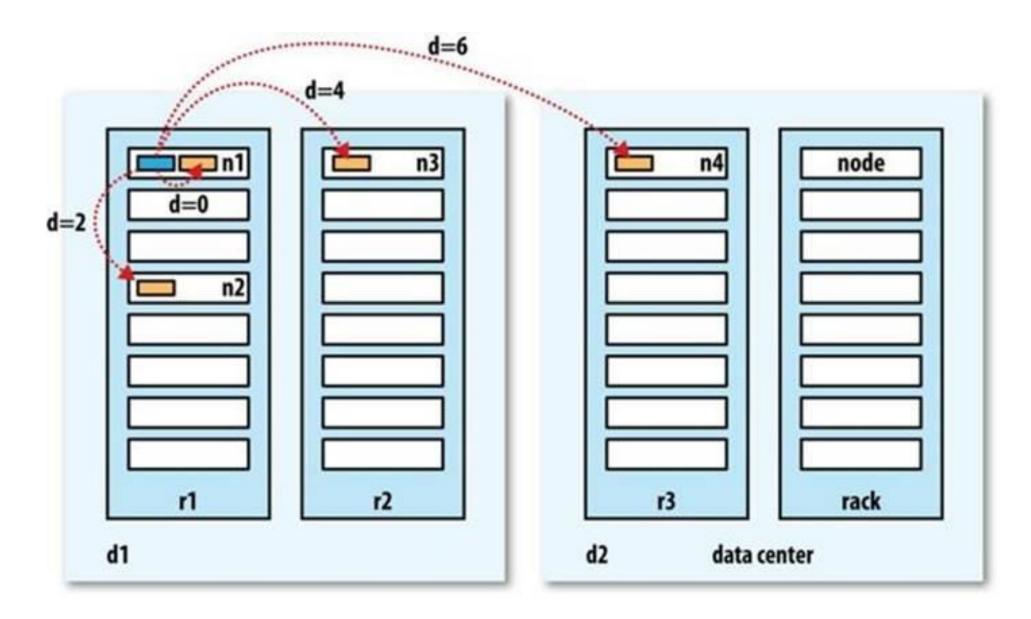


Hub & Spoke Hardware



- Typically in 2 level architecture
 - Nodes are commodity PCs
 - Typically 30-40 nodes/rack

Notion of Distance [wrt data transfer]



Hadoop Characteristics:

Distribute data initially

- Let processors / nodes work on local data
- Minimize data transfer over network
- Replicate data multiple times for increased availability

Write applications at a high level

- Programmers should not have to worry about network programming, low level infrastructure, etc
- Minimize talking between nodes (share-nothing)

Requirements Speed Scalability Reliability Ease of programming Low cost

Why Is Hadoop Important?

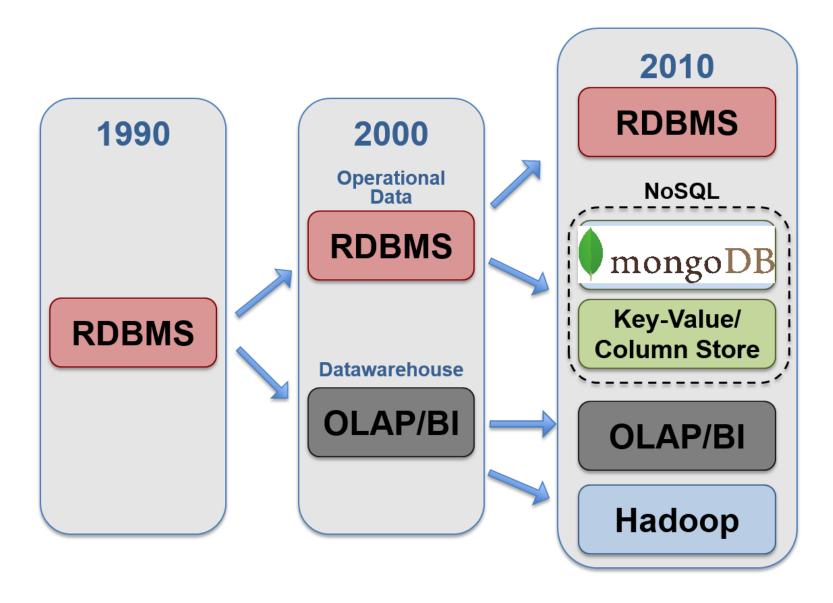
- Ability to store and process huge amounts of any kind of data, quickly.
- Computing model processes big data fast
- Fault tolerance
- Flexibility
- Low Cost
- Scalability

Break

Let's set the stage for next discussion:

- RDBMS
- SQL
- JDBC
- OLTP
- OLAP
- NoSQL
- Data Warehouse
- ETL
- Data Mart
- Data Lake

Evolution of Databases







- All data has potential value
- Data hoarding
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time (schema-on-read).
- Apps and users interpret the data as they see fit

The "Big Data Bazaar"

Key Hadoop Vendors [By Type]

- Pure play Hadoop vendors: Cloudera + Hortonworks, MapR, IBM OpenPlatform, Huawei FusionInsight, Seabox, Transwarp
- Cloud infrastructure as a service (laaS): Hadoop on AWS, Hadoop on Azure
- Platform as a service (PaaS): IBM BigInsights, Microsoft HDInsight, Google Cloud Platform, Amazon EMR, Oracle Big Data Cloud Service, Qubole
- Big Data Appliances: Teradata, Oracle Big Data Appliance, Cray

Gartner.

Market Guide for Hadoop Distributions

Published: 01 February 2017 ID: G00298214

Analyst(s): Nick Heudecker, Merv Adrian, Ankush Jain

Market Position of Hadoop Vendors

Gartner Magic Quadrant for Data Management Solutions for Analytics (DMSA) - 2017

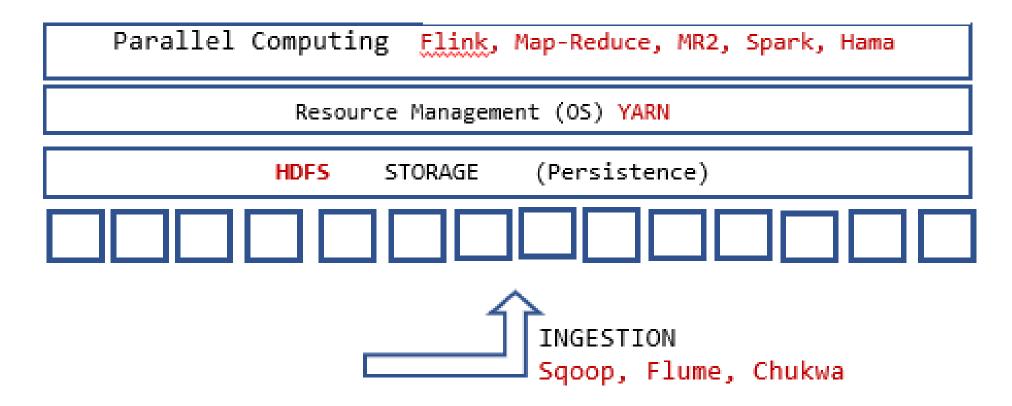


As of February 2057

Big data ecosystem

- Programmers ought to see the storage as monolithic.
 - Resource Management: YARN, Mesos
- "Serially written" programs ought to run in parallel.
 - Map Reduce, MR2, Spark, BSP, Flink, ...
- There ought to be faster, more reliable ways of bringing in much more data
 - Data ingestion methods Sqoop, Flume, Kafka...

Everyone needs them.....



Computing Layer

Machine Learning on Spark-ML, Mahout, Samsara, H20, Hadoop

Streaming & Near Real Time Processing

KAFKA, SAMZA, STORM,
TRIDENT, SPARK-STREAMING,
FLINK

Pick what you need

Application Programming

PIG, Oozie, Hadoop Streaming, Spark-R

Data Organization
SQL
HIVE, IMPALA, SPARK SQL, Apache Drill
NoSQL
Hbase, Cassandra, MongoDB, Neo4J, Kudu

Security, Audit, Governance

```
SECURITY
&
QOS
```

KNOX Ranger Sentry Atlas Kerberos

PRIVACY

AUDIT

GOVERNANCE

Primary Content of Hadoop Distributions: Apache Projects

All major Hadoop vendors support these Apache projects:

HDFS, MapReduce, YARN, Pig, Hive, Hbase, ZooKeeper, Avro, Flume, Kafka, Oozie, Parquet, Solr, Spark and Sqoop.

Select vendors support these Apache projects:

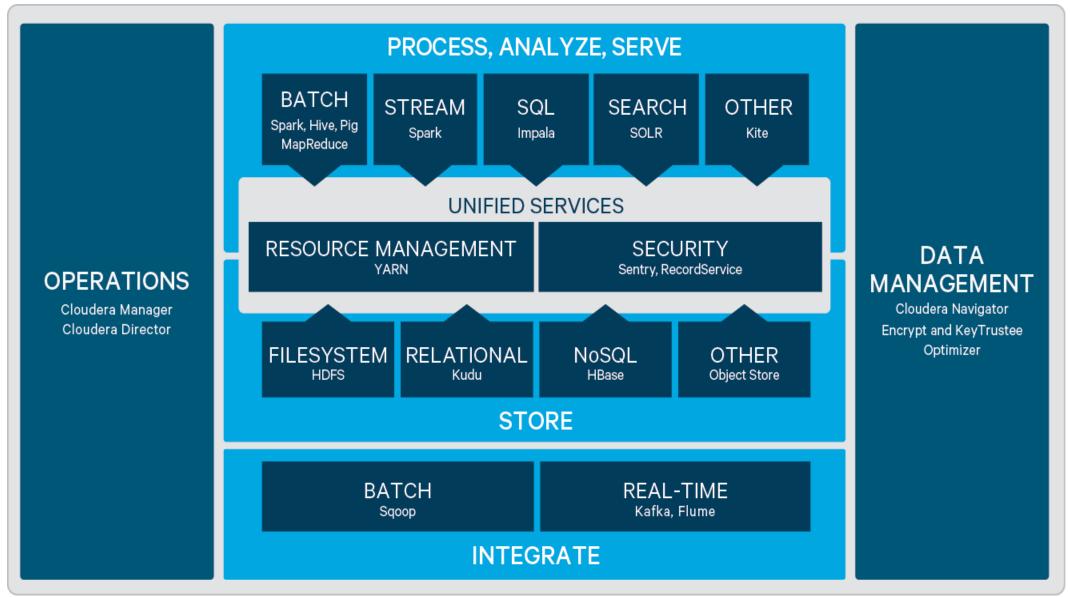
Accumulo, Ambari, Atlas, Impala, Knox, Myriad, NiFi, Phoenix, Slider, and Zeppelin

These projects show great promise:

Apache Arrow, Beam, Flink, Ignite and Kylin

Details & More: https://hadoopecosystemtable.github.io/

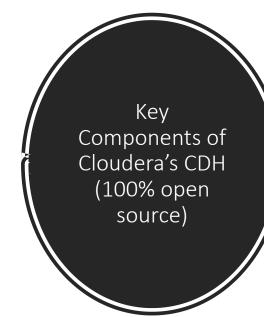
Sample Hadoop Distribution: Cloudera





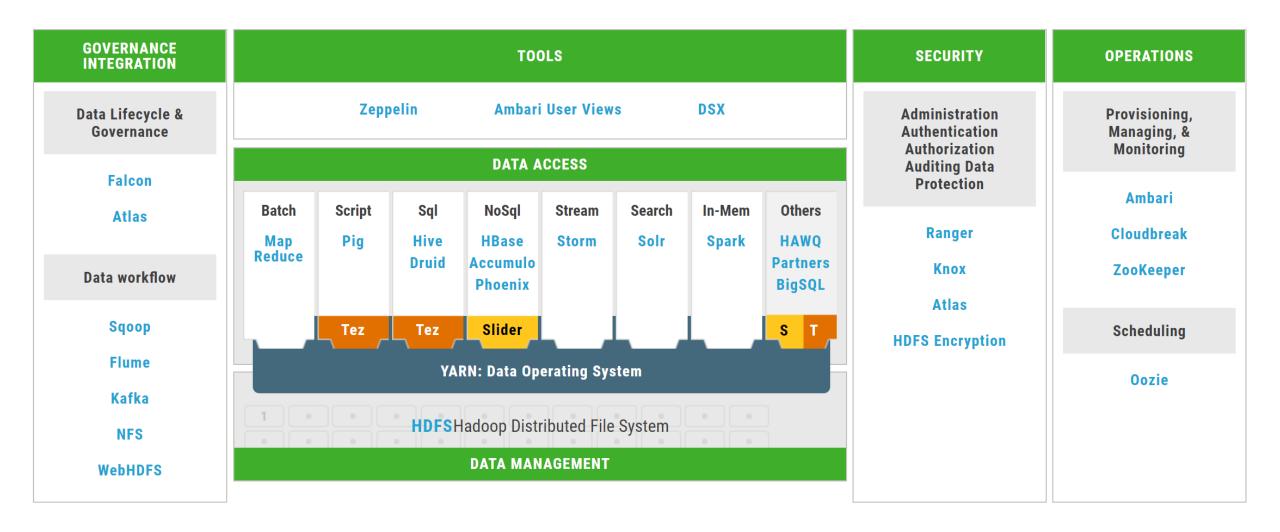


- Pig
- Kafka
- Spark
- Sqoop, Flume
- Impala
- Accumulo

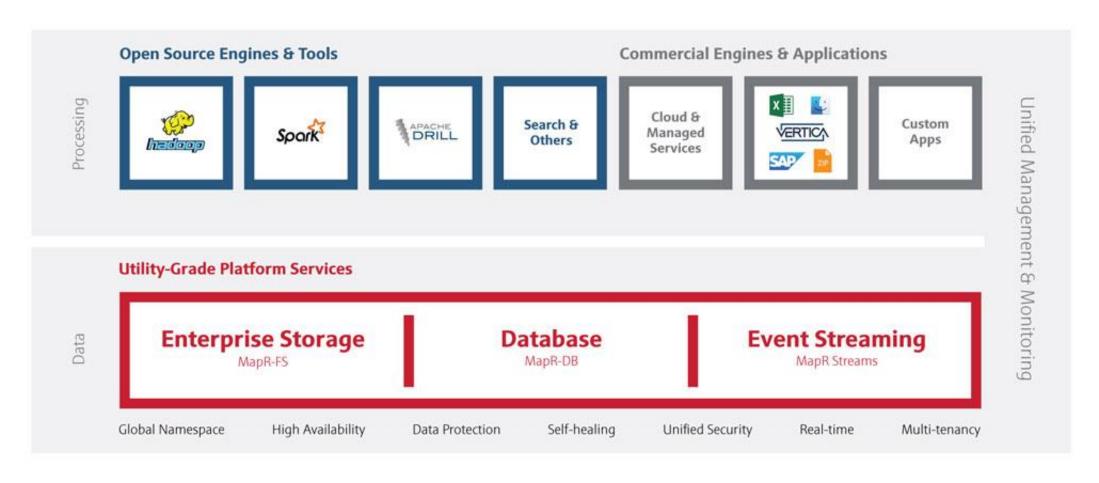




Hortonworks



MapR Converged Data Platform



Big Data services on AWS

Collect







Store











Amazon RDS,

Aurora

Amazon

Elasticsearch

Service

Process & Analyze







Amazon EMR





Amazon QuickSight



Amazon Redshift



Amazon Machine Learning



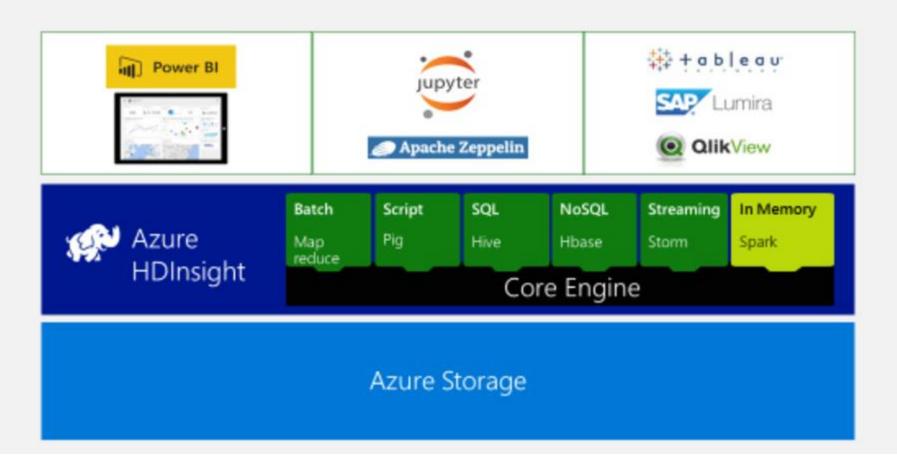
Amazon Kinesis Analytics

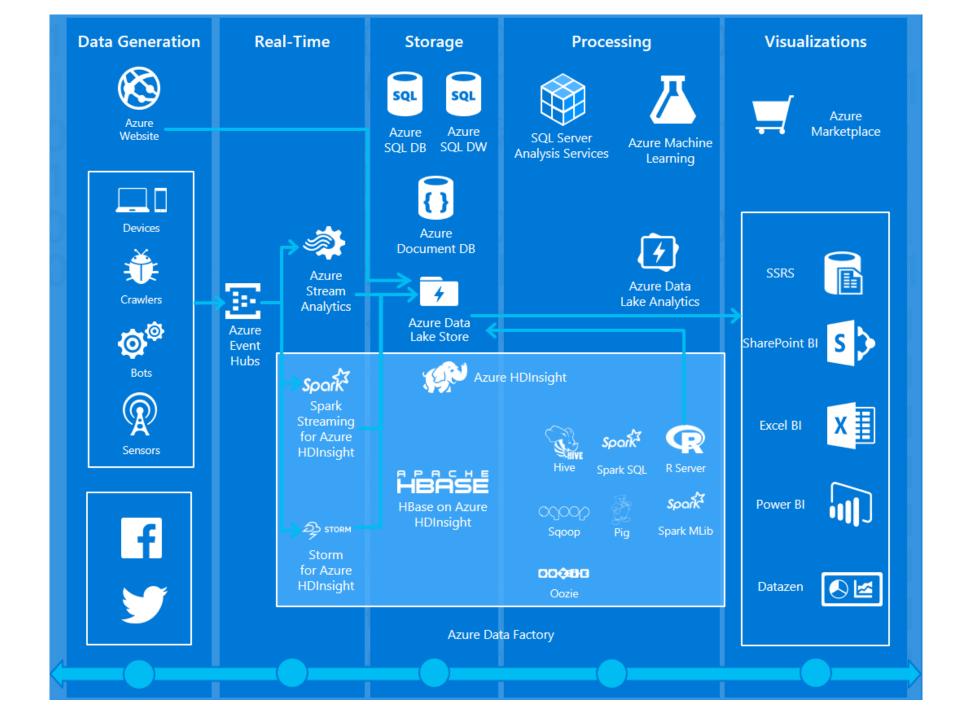


Spark For Azure HDInsight

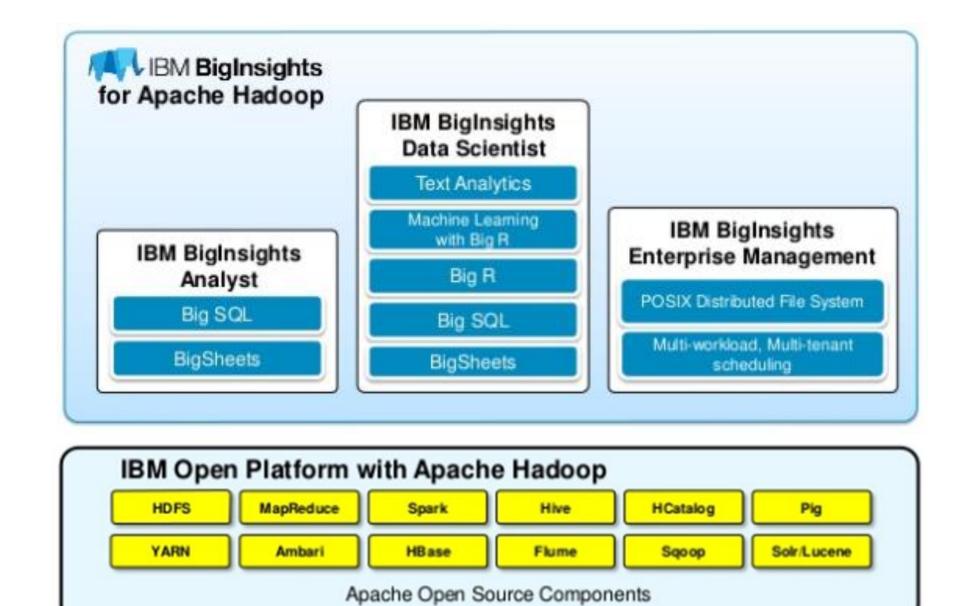


Choice of compelling interactive experiences

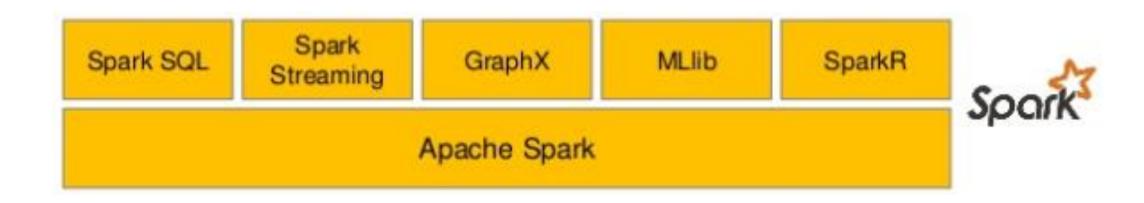




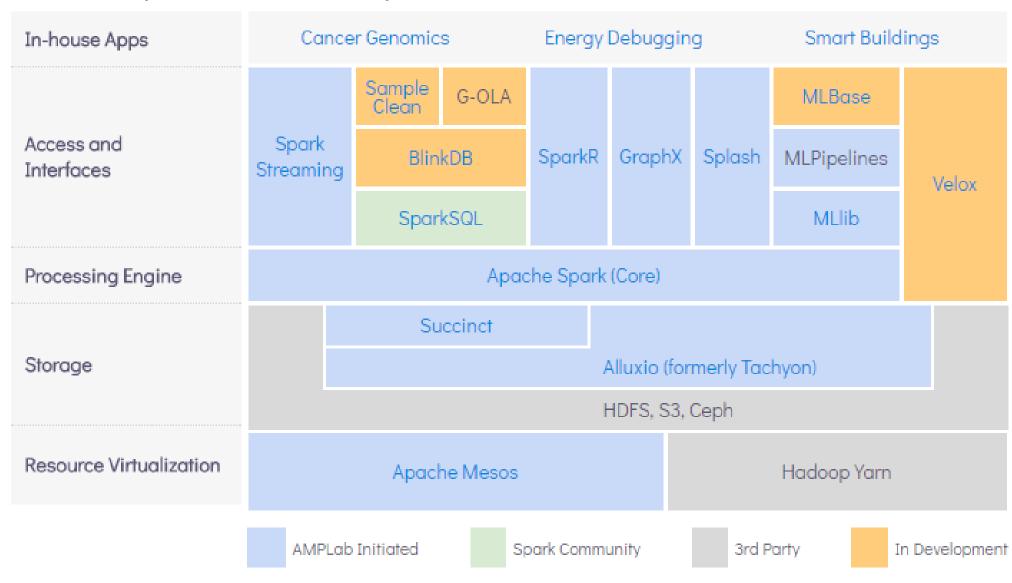
Microsoft Azure HDInsight Architecture



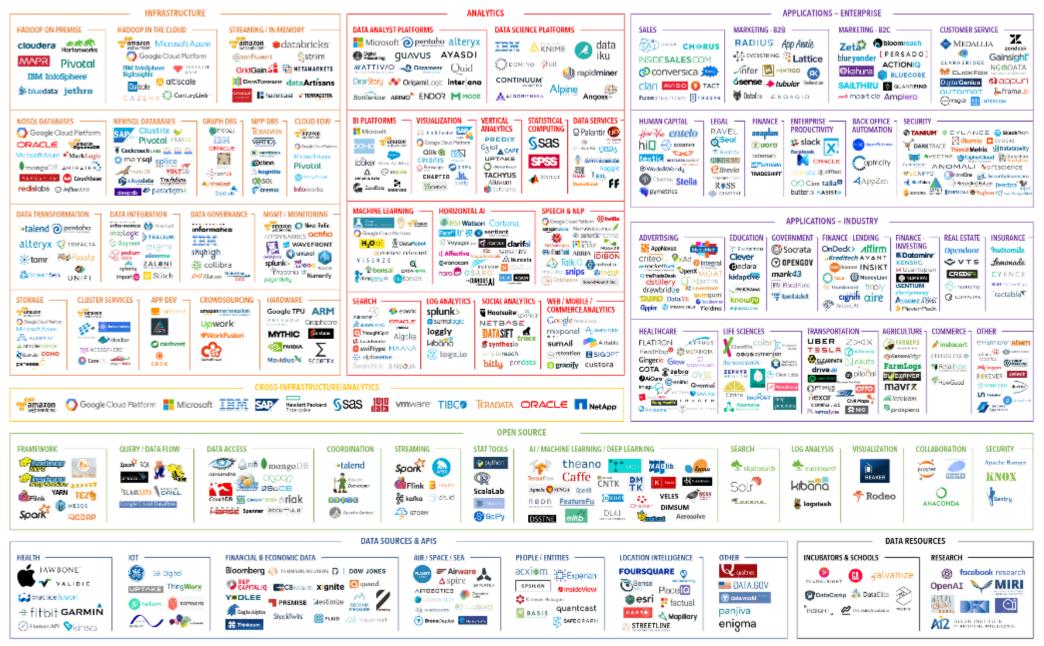
The Spark Ecosystem



Berkeley Data Analytics Stack (BDAS)



BIG DATA LANDSCAPE 2017



V2 - Last updated 5/3/2017

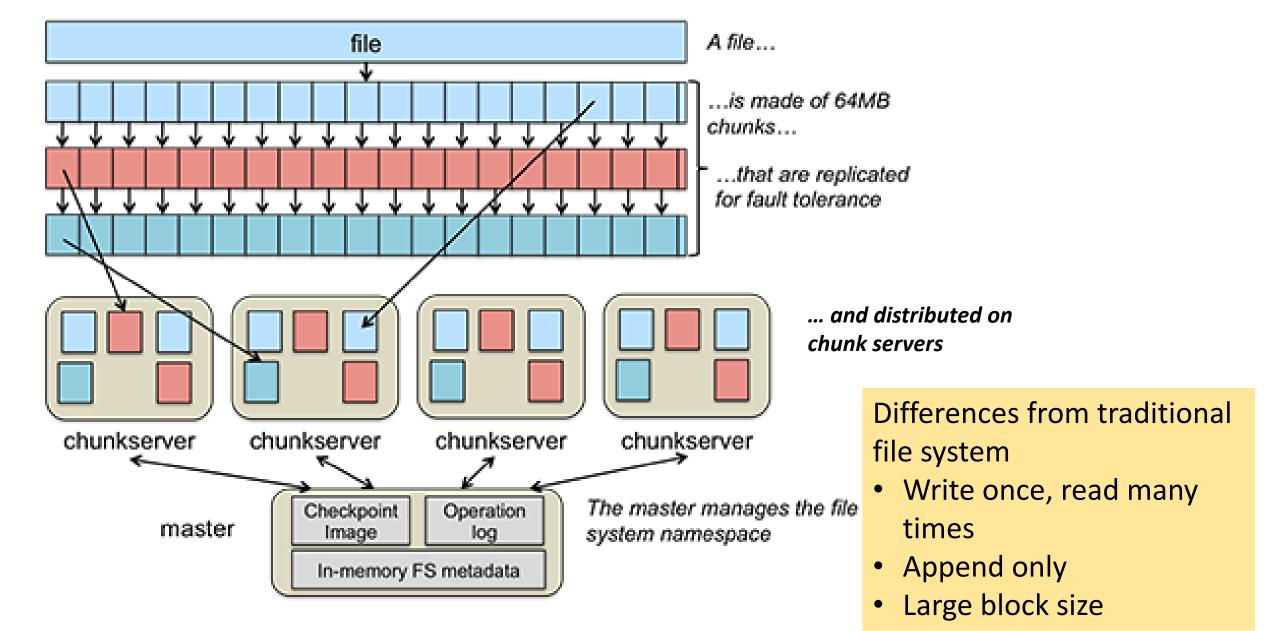


"So you want to hire me as a Data Scientist for Intelligent Virtualized Deep Machine Learning Real-time Big Data in the Cloud for Social Networks? Ok, but if you also want Hadoop, increase my salary by 50%."

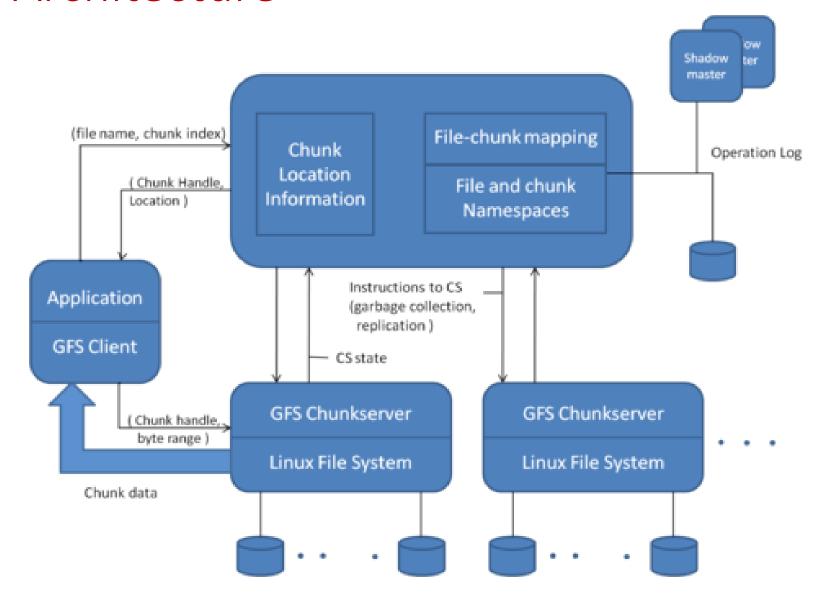
Hadoop: Storage

HDFS

Background: Google File System (GFS)



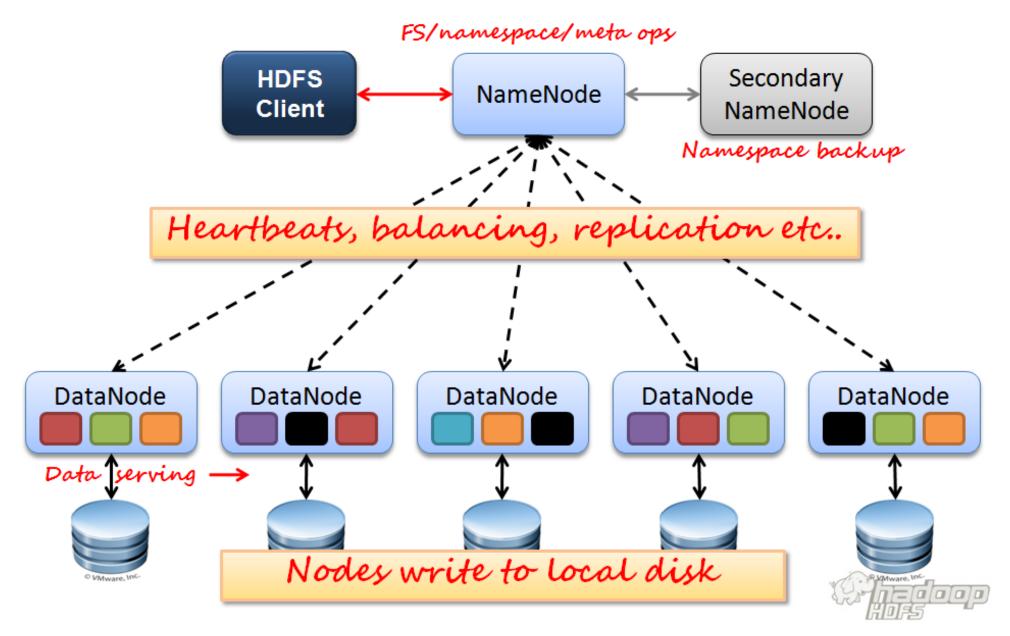
GFS: Architecture



GFS Master Responsibilities

- Metadata storage
- Namespace management/locking
- Periodic communication with chunkservers
 - □ give instructions, collect state, track cluster health
- Chunk creation, re-replication, rebalancing
 - □ balance space utilization and access speed
 - □ spread replicas across racks to reduce correlated failures
 - □ re-replicate data if redundancy falls below threshold
 - □ rebalance data to smooth out storage and request load
- Garbage Collection
 - □ simpler, more reliable than traditional file delete
 - ☐ master logs the deletion, renames the file to a hidden name
 - □ lazily garbage collects hidden files
- Stale replica deletion
 - □ detect "stale" replicas using chunk version numbers

HDFS CDH3: Open source reimplementation of GFS



NameNode Metadata

- Metadata in Memory
 - The entire metadata is in main memory

- Types of metadata
 - List of files
 - List of Blocks for each file
 - List of Data Nodes for each block
 - File attributes, e.g. creation time, replication factor
- A Transaction Log
 - Records file creations, file deletions etc

Block Replica Placement

- Current Strategy
 - One replica on local node
 - Second replica on a remote rack
 - Third replica on same remote rack
 - Additional replicas are randomly placed
- Clients read from nearest replicas

Policy is pluggable

Replica Block Placement

Rack-Aware strategy

HDFS: Heartbeat and Rebalancing

Heart beats

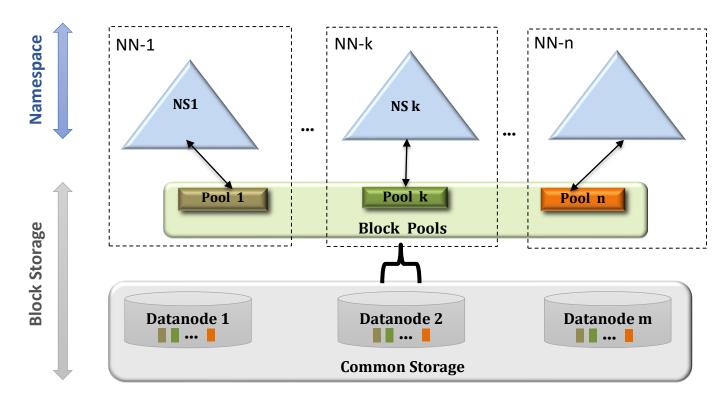
- Data Nodes send heart beat to the Name Node
- Once every 3 seconds
- Name Node uses heartbeats to detect Data Node failure

- Rebalancing: % disk full on Data Nodes should be similar
 - Usually run when new Data Nodes are added
 - Cluster is online when Rebalancer is active
 - Rebalancer is throttled to avoid network congestion
 - Command line tool

What limitations you can think of?

Any problems you foresee wrt what we have seen so far?

HDFS 2.0: Name Node Federation Elaborated



- Multiple independent Namenodes and Namespace Volumes in a cluster
 - Namespace Volume = Namespace + Block Pool
- Block Storage as generic storage service
 - Set of blocks for a Namespace Volume is called a *Block Pool*
 - DNs store blocks for all the Namespace Volumes no partitioning

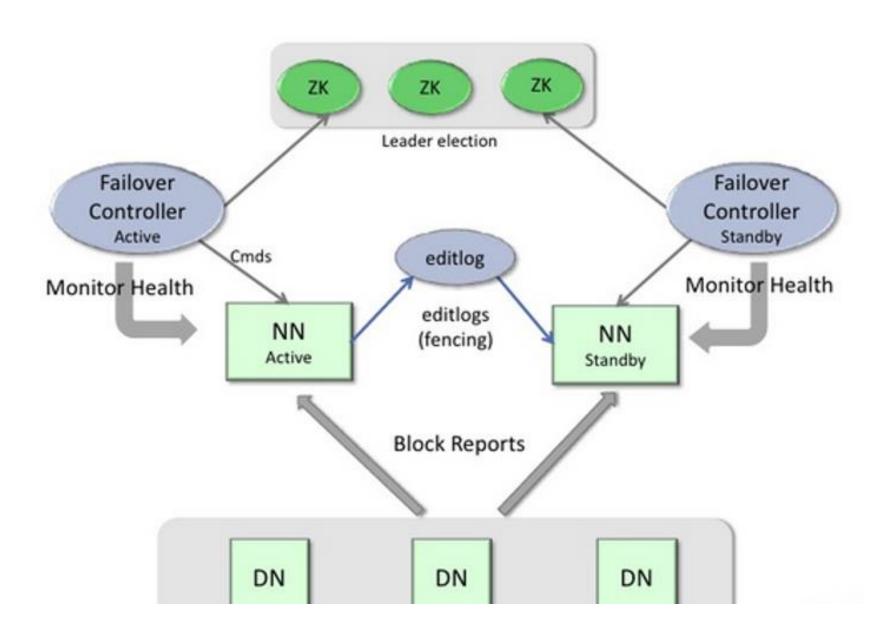
What benefits come from "Federation"?

Bottlenecks

Chargeback

Isolation

HDFS 2.0: High Availability Elaborated



In order to provide a fast failover, it is also necessary that the Standby node have up-to-date information regarding the location of blocks in the cluster.

In order to achieve this, the DataNodes are configured with the location of both NameNodes, and send block location information and heartbeats to both.

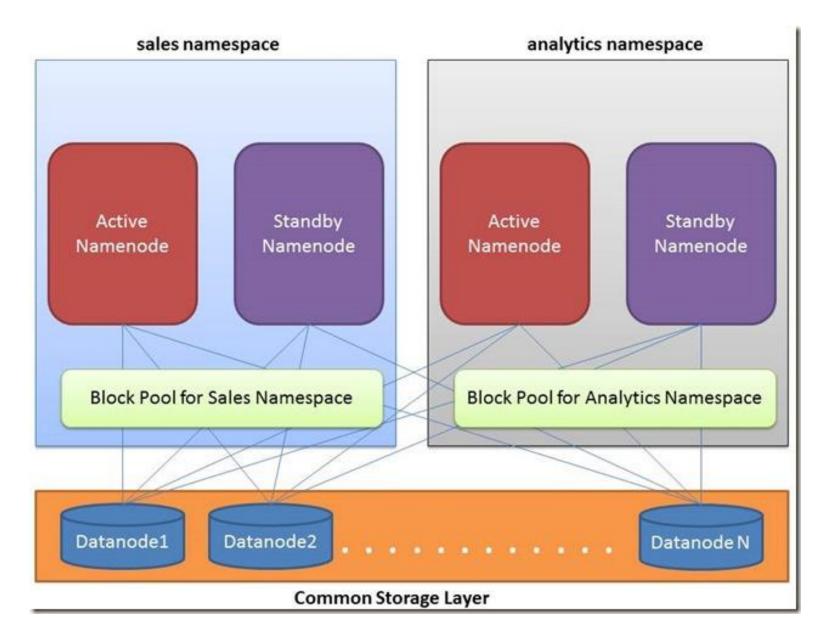
The ZKFailoverController (ZKFC) is a ZooKeeper client that also monitors and manages the state of the NameNode.

Each of the hosts that run a NameNode also run a ZKFC.

The ZKFC is responsible for **Health monitoring** of Namenode

- the ZKFC contacts its local NameNode on a periodic basis with a health-check command. So long as the NameNode responds promptly with a healthy status, the ZKFC considers the NameNode healthy.
- If the NameNode has crashed, frozen, or otherwise entered an unhealthy state, the health monitor marks it as unhealthy.

HDFS 2.0: High Availability, Federated



Some administrative work

Block Checker

Balancer

Commissioning and decommissioning

- HDFS still needs to be backed up.
- Select and Prioritize what to backup

So what all we have learnt?

V's of Big Data

• Big Data Use Cases

Hadoop and its ecosystem

• HDFS, HDFS 2.0

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-

school-of-engineering

Thank you! prasadmd0@gmail.com