Trendwise Analytics

Big Data/Hadoop Introduction

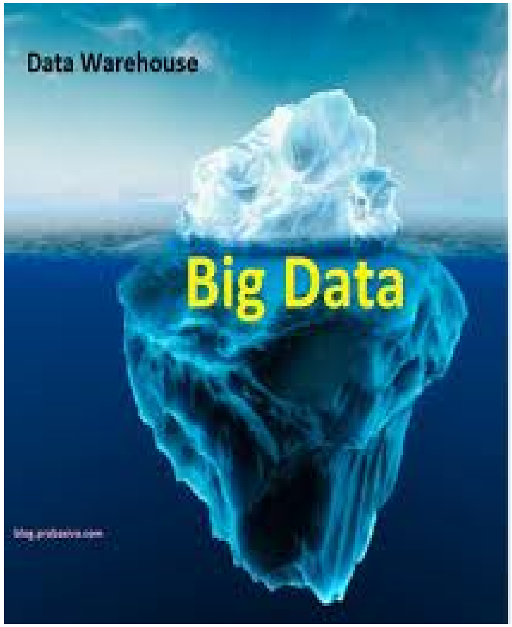
Agenda

1. Introduction to Big Data and Hadoop
2. Big Data Business cases
3. Technology
4. Q&A - Wind up

What is Big Data

Three V's

- **Volume**
- **Variety**
- **Velocity**



Volume of Data

'Black Box'
Flight Data Recorder
Power supply
Crash Survivable Memory Unit (CSMU)
Underwater locator beacon
Connection to aircraft
Controller board
49.7cm

- A commercial aircraft generates 3GB of flight sensor data in 1 hour

An ERP system for an mid size company grows by 1-2TB annually

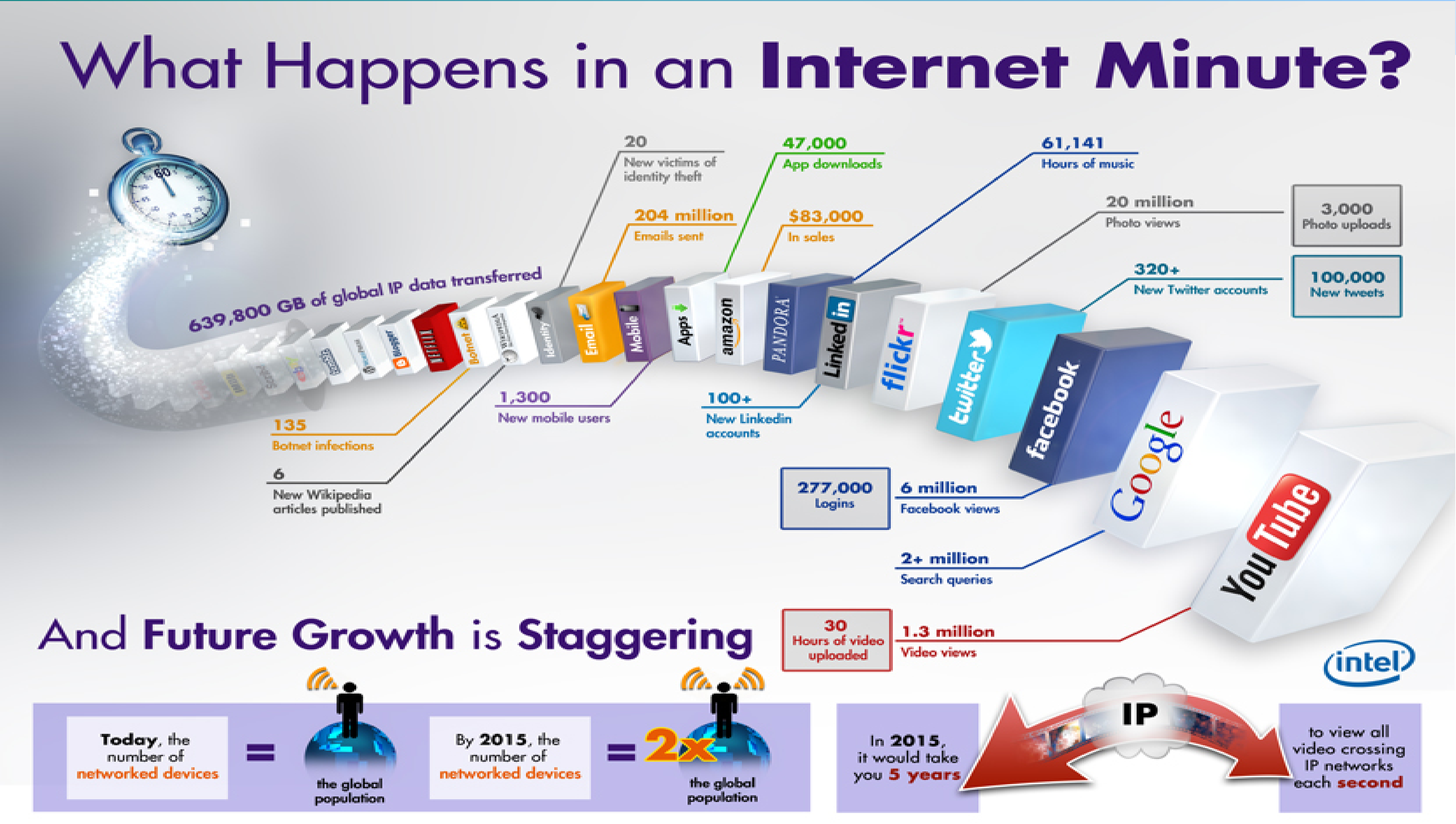
A Video Suveillance Camera generates 1-3TB data in 3 months

Airtel or Vodafone generates 3TB of Call Details Records (CDR) every day

Every day 2.5 quintillion (2.5×10^18) bytes of data is created

i.e., 2,500,000TB

Internet Minute ....



What Happens in an **Internet Minute?**

639,800 GB of global IP data transferred

20
New victims of identity theft

204 million
Emails sent

47,000
App downloads

$83,000
In sales

61,141
Hours of music

20 million
Photo views

3,000
Photo uploads

320+
New Twitter accounts

100,000
New tweets

135
Botnet infections

6
New Wikipedia articles published

1,300
New mobile users

100+
New Linkedin accounts

277,000
Logins

6 million
Facebook views

2+ million
Search queries

30
Hours of video uploaded

1.3 million
Video views

And **Future Growth** is **Staggering**

Today, the number of networked devices = the global population

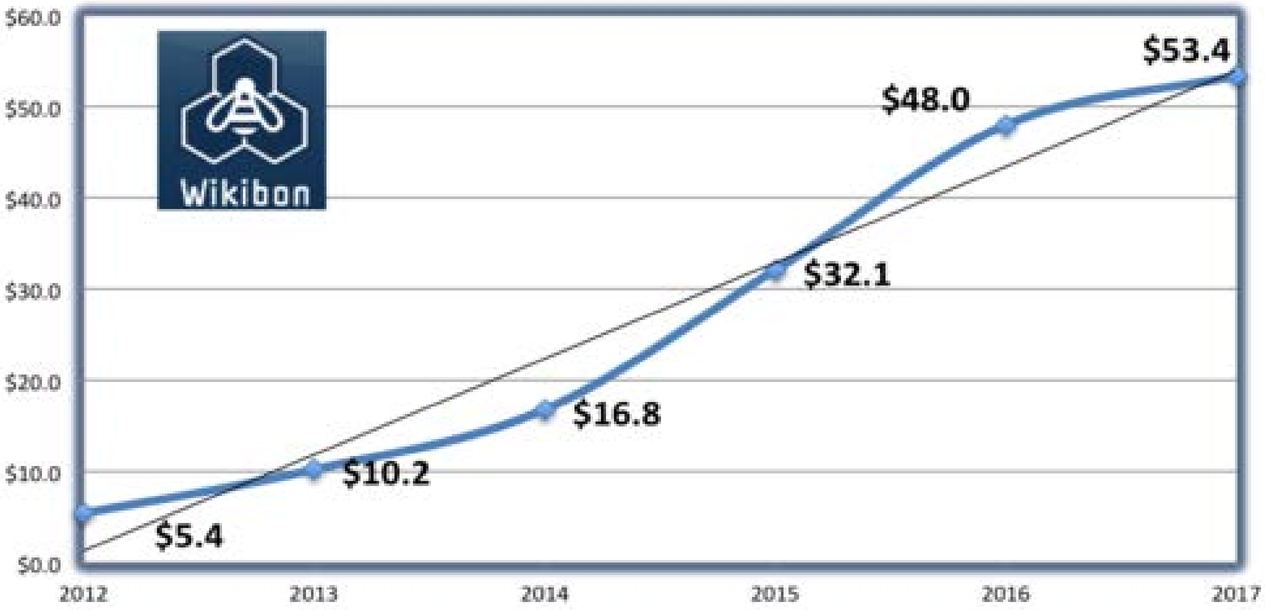
By **2015**, the number of networked devices = 2x the global population

In **2015**, it would take you **5 years** to view all video crossing IP networks each **second**

Market opportunity

IDC, a research firm, predicts that the market for Big Data technology and services will reach $16.9 billion by 2015, up from $3.2 billion in 2010. That is a 40 percent-a-year growth rate — about seven times the estimated growth rate for the overall information technology and communications business, according to IDC.

Big Data Market Forecast, 2012-2017 (in $US billions)



Billions and billions: big data becomes a big deal :

Deloitte predicts that in 2012, “big data” will likely experience accelerating growth and market penetration.

How Companies are using Big Data?

Common Big Data Customer Scenarios in your industry

IT infrastructure optimization

Legal discovery

Social network analysis

Traffic flow optimization

Web app optimization

Churn analysis

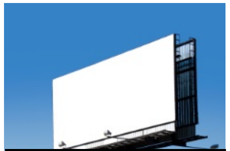
Natural resource exploration

Weather forecasting

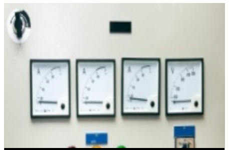
Healthcare outcomes

Fraud detection

Life sciences research



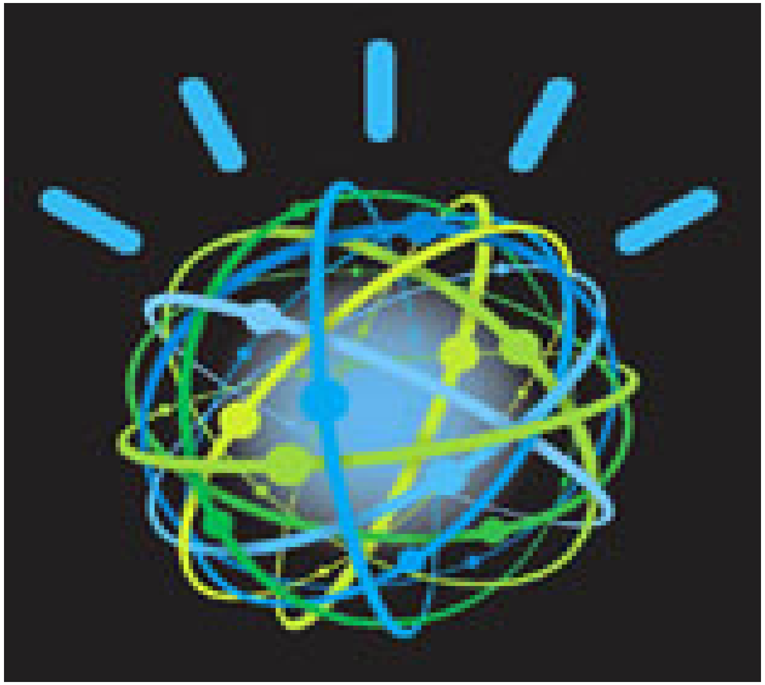
Advertising analysis



Equipment monitoring

Smart meter monitoring

Watson wins Jeopardy!



- Feb 14th 2011 – Watson wins Jeopardy! beating its human opponents.
- Watson is IBM's super computer built using Big Data Technology.

Big data Applications

- Social media analytics – ”People You May Know” at LinkedIn
- Voice analytics – Call center
- Text analytics – Voice of customer, sentiment analysis, warranty analysis
- Video analytics – Intelligence, policing, retail applications
- Telecom – customer churn

Big Data at GE

- **New $1B corporate center for software and analytics**
  - Hiring 400 data scientists
- **Includes financial and marketing applications,**

  but with special focus on industrial uses of big data
  - When will this gas turbine need maintenance?
  - How can we optimize the performance of a locomotive?
  - What is the best way to make decisions about energy finance?

Ford Gets Smarter About Marketing and Design

- Ford collects and aggregates data from the **4 million** vehicles that use in-car sensing and remote app management software
- The data allows to glean information on a range of issues, from how drivers are using their vehicles, to the driving environment that could help them improve the quality of the vehicle
- Partnered with Microsoft to develop SYNC

How Amazon Uses Big Data To Make You Love Them

- Amazon has been collecting customer information for years--not just addresses and payment information but the identity of everything that a customer had ever bought or even looked at.
- They’re using that data to build customer relationship

How LinkedIn is Riding a Wave of Big Data All the Way to the Bank

- LinkedIn is a trove of data not just about people, but how people are making their money and what industries they are working in and how they connect to each other.

How AT&T is using cell phone to watch user movements?

- AT&T has **300 million** customers
- A team of researchers is working to turn data collected through the company’s cellular network into a trove of information for policymakers, urban planners and traffic engineers.
- The researchers want to see how the city changes hourly by looking at calls and text messages relayed through cell towers around the region, noting that certain towers see more activity at different times

Govt of India

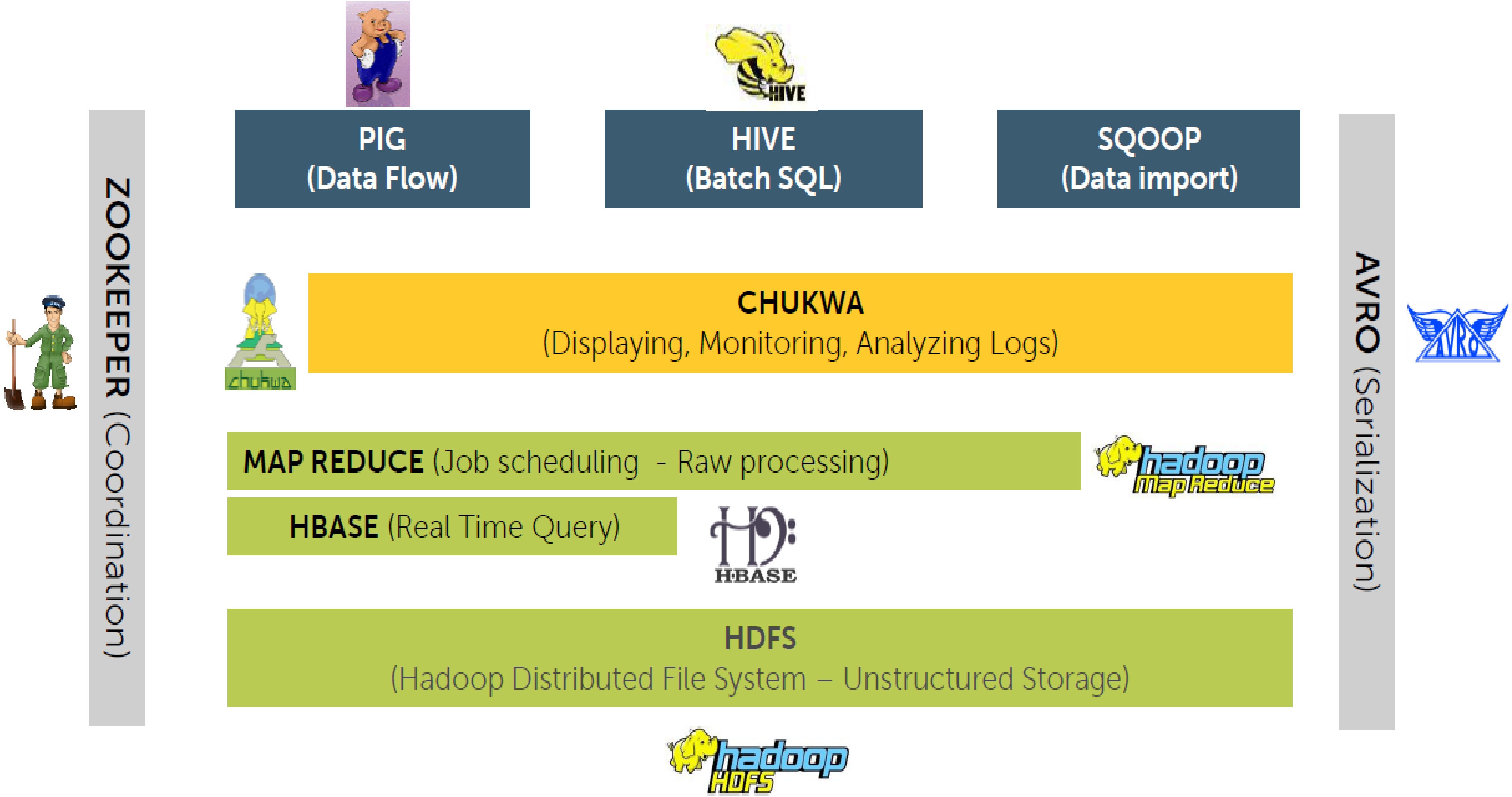
- Aadhar project by Govt. of India uses Hadoop

TECHNOLOGY

Hadoop

Non Hadoop

Hadoop Components



PIG
(Data Flow)

HIVE
(Batch SQL)

SQOOP
(Data import)

CHUKWA
(Displaying, Monitoring, Analyzing Logs)

MAP REDUCE (Job scheduling - Raw processing)

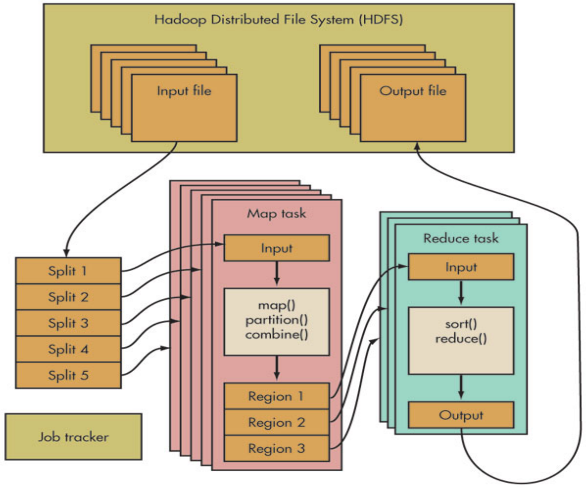
HBASE (Real Time Query)

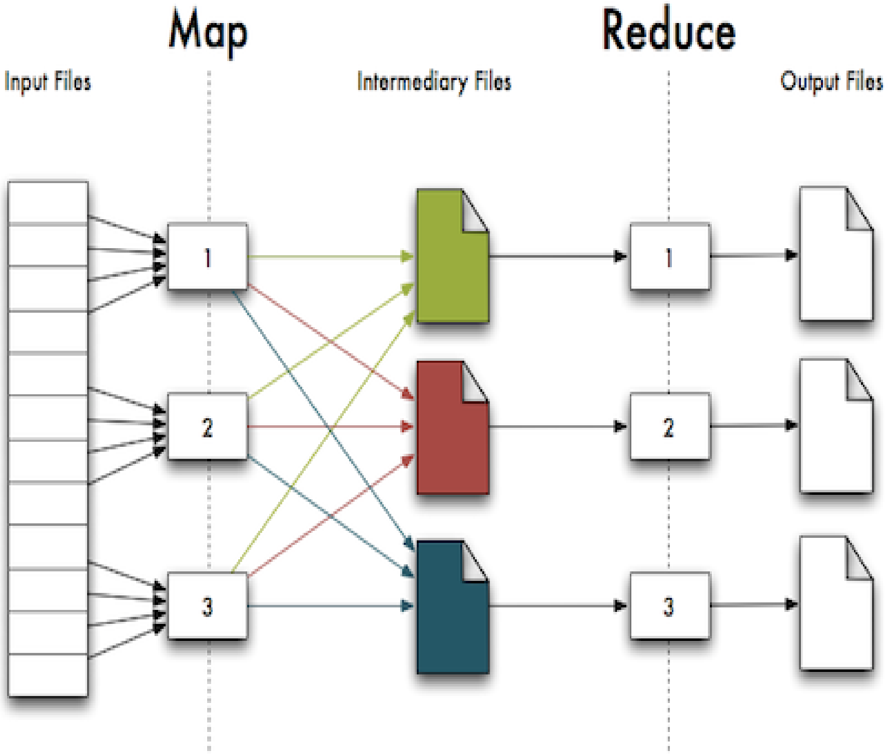
HDFS
(Hadoop Distributed File System – Unstructured Storage)

ZOOKEEPER (Coordination)

AVRO (Serialization)

Hadoop HDFS and MapReduce





- Hadoop runs on HDFS,
  Hadoop Distributed Filesystem
- Any data stored is converted to blocks
  and distributed across the cluster nodes

Other components

Hive

- Data Warehouse infrastructure that provides data summarization and ad hoc querying on top of Hadoop

Sqoop

- Sqoop is a tool designed to help users of large data import existing relational databases into their Hadoop clusters

PIG

- A high-level data-flow language and execution framework for parallel computation

Zookeeper

- Zookeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services

Benefits of Hadoop

- Hadoop is designed to run on cheap commodity hardware
- It automatically handles data replication and node failure
- Handles large volumes of unstructured data easily
- Last but not least – its free! (Open source)

Commercial Hadoop Distributions

- Cloudera
- Hortonworks
- Greenplum, A Division of EMC
- IBM InfoSphere BigInsights

Technology – Non Hadoop

- **HPCC** - HPCC Systems from LexisNexis Risk Solutions offers a proven,
  - open-source, data-intensive supercomputing platform
    - designed for the enterprise to solve big data problems.

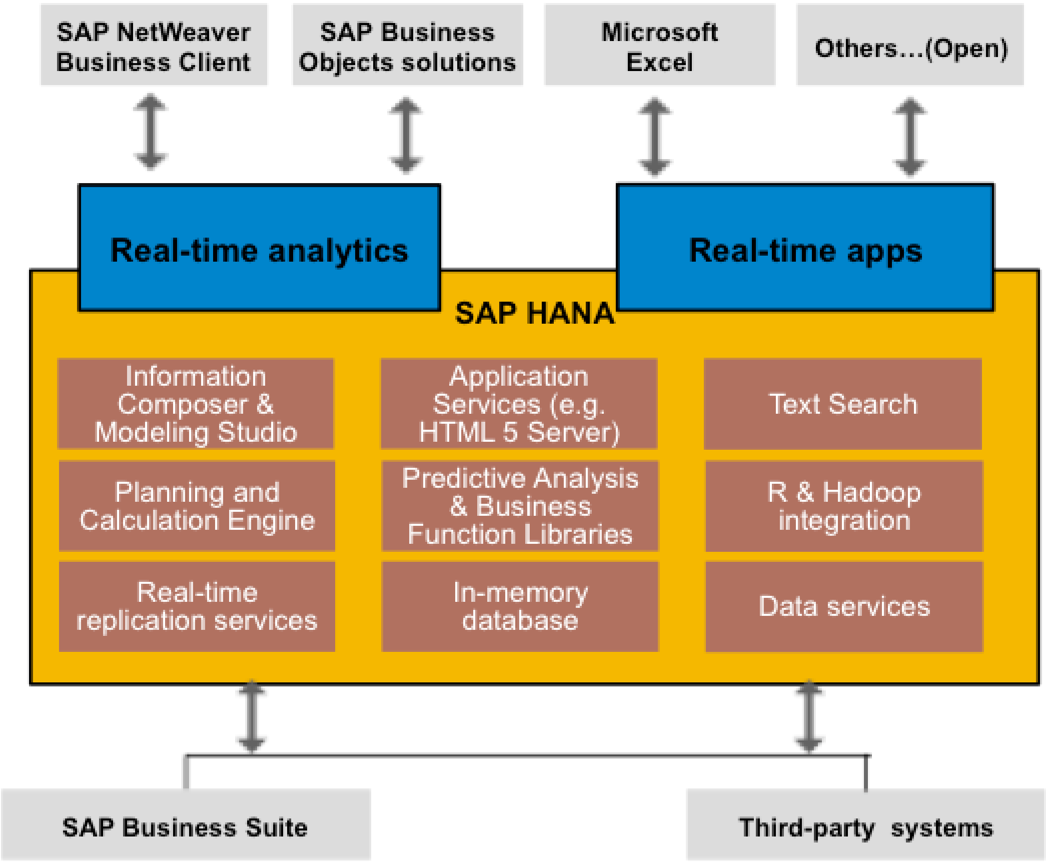
- **SAP HANA** is SAP AG’s implementation of in-memory database technology.

- **NoSQL Databases**

  Key-Values Stores – Redis, Riak
  Column Family Stores – Cassandra, HBase
  Document Databases – CouchDB, MongoDB
  Graph Database – InfoGrid, Infinite Graph

SAP HANA In-memory Database System

- Hana is an in-memory database system developed by SAP AG.
- It takes the advantage of –
  - low-cost of main memory
  - Fast data access of solid state drives.
  - Data processing abilities of multi-core processors.
- It supports both row-oriented and column-oriented data storage.
- It incorporates powerful graph and text processing capabilities to work with semi and full unstructured data.
- SAP has positioned HANA as its solution to big data challenges at the low end of this scale.


SAP NetWeaver Business Client
SAP Business Objects solutions
Microsoft Excel
Others...(Open)
Real-time analytics
Real-time apps
SAP HANA
Information Composer & Modeling Studio
Application Services (e.g. HTML 5 Server)
Text Search
Planning and Calculation Engine
Predictive Analysis & Business Function Libraries
R & Hadoop integration
Real-time replication services
In-memory database
Data services
SAP Business Suite
Third-party systems

Thank You!

Additional resources

Hadoop:

http://hadoop.apache.org/

http://bigdatauniversity.com/

Others:

http://www.cloudera.com/content/cloudera/en/home.html

http://hortonworks.com/