HADOOP & BIG DATA

Senthil Kumar A

DATA DOTZ

JPA Solutions
Center Of Excellence

About Me

- Senior Solution Architect (BigData) at USEReady
- Chief Technical Advisor to DataDotZ
  - DataDotZ – BigData Training Partner for JPA Solutions
- Technical Speaker
  - Anna University, VIT University, KSR College of Engineering.
- Founding Member of Chennai Hadoop Users Group
  - *https://groups.google.com/group/chennaihug*

Agenda

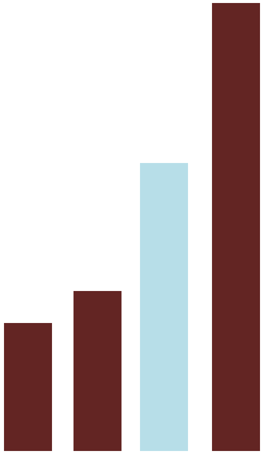
- What is Big Data??
- What is Hadoop EcoSystem??
- Relationship between Hadoop and BigData

~~Big~~ Data

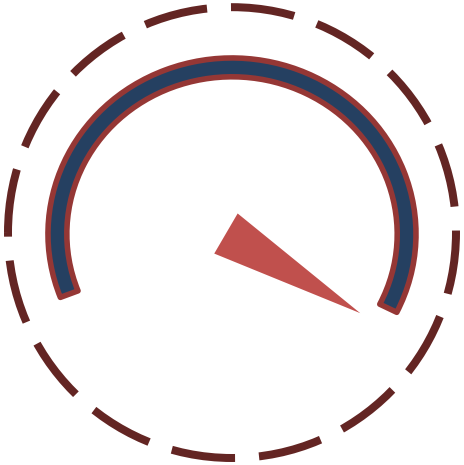
- Storage
  - Flat Files
  - RDBMS
  - InMemory DataGrid
  - NAS,SAN
  - NoSQL
- Computation
  - *

Big Data

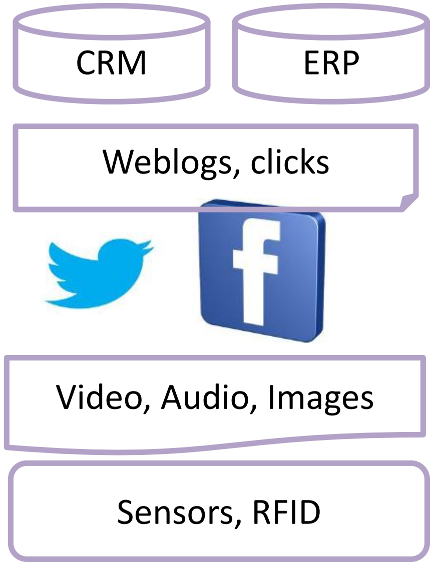
Storage , Computation



Volume



Velocity



CRM

ERP

Weblogs, clicks

Video, Audio, Images

Sensors, RFID

Variety

Structured Data
Semi Structured Data
UnStructured Data

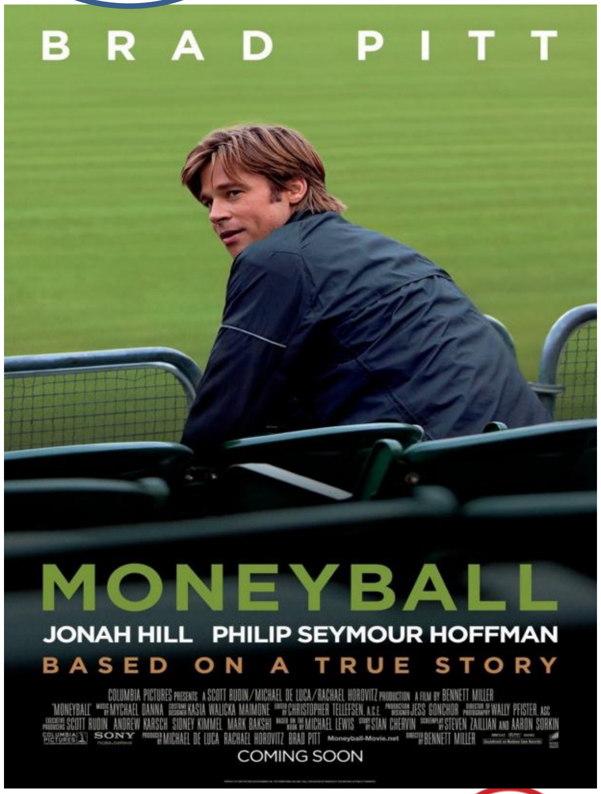
Some Companies

- IBM – 4 Vs – 4th V is Veracity
- SAS – 4th – Variability
- Microsoft (& Others) – 4th - Value

Definition of BigData

- In our terms , BigData is a problem statement
- The problems may be arised due to
  - Heavy Storage
  - Heavy Computation
  - Both

Data Driven Decisions



Traditional Systems

- Small Amount of data - RDBMS
  - Less Data -> Less IO
  - Performance based on processor as well as RAM
- Scalability
  - Sharding – RDBMS
- Proprietary Systems
- Distributed Storage
  - SAN , NoSQL
  - What about Computation? Lots of IO?

FlashBack

- 2002 - Nutch for web crawling & search
  - Doug Cutting & Mike Cafarella
- 2003 - Google published GFS paper
- 2004 - NDFS
- 2004 - Google published MapReduce paper
- 2005 - Mapreduce + NDFS
- 2006 - Formed Subproject (HADOOP)
- 2006 - Doug joins Yahoo
- 2008 - World record (Terasort)

Hadoop EcoSystem

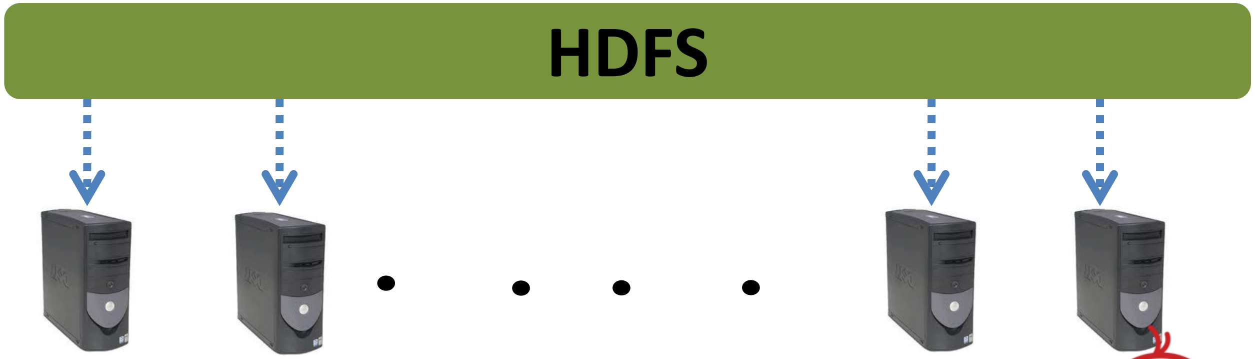
Not a Tool, It’s a Framework !!!

- Distributed System
  - Storage and Computation
- Reduce the IO
  - Move the Computation to Data (Grey’s Third Law) - Data Locality
- Data Recoverability
  - Fault Tolerant System
- Scalability and Performance
  - Scale Out Architecture
  - Linear increase in performance
- Open Source
  - Support available
- Commodity Servers

Hadoop Distributed File System

Distributed Storage

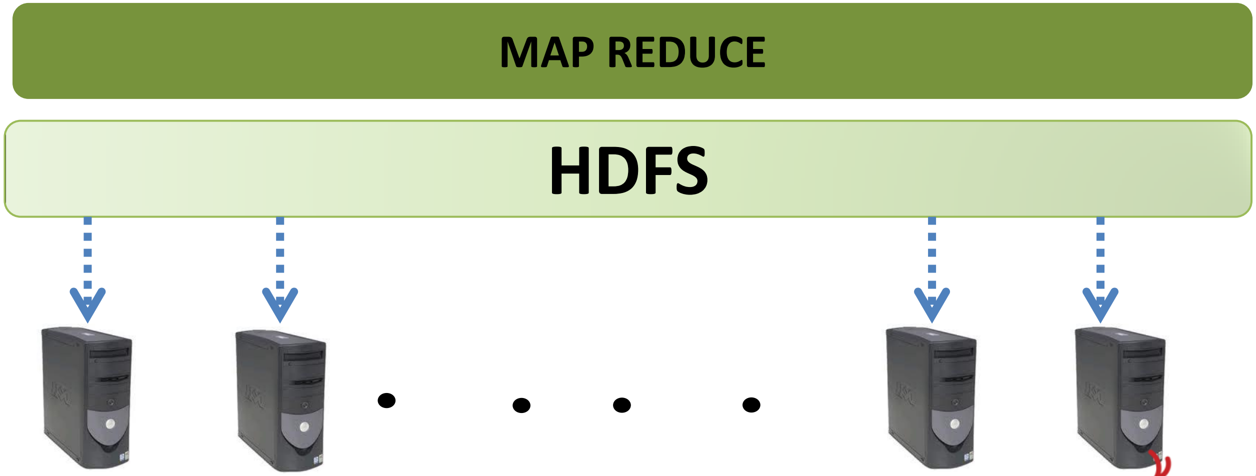
- Concept of Blocks
- Stores using Local FileSystem
- Fault Tolerant by replication
- Data Pipelining
- Coherency
- Distributed across Machines



MapReduce

Distributed Computation

- Distributed Parallel Processing
- Data Locality
- Codesigned , colocated , codeployed with HDFS
- Complete Abstraction – Programming APIs exposed
- Component Failure Recovery
- Consistency



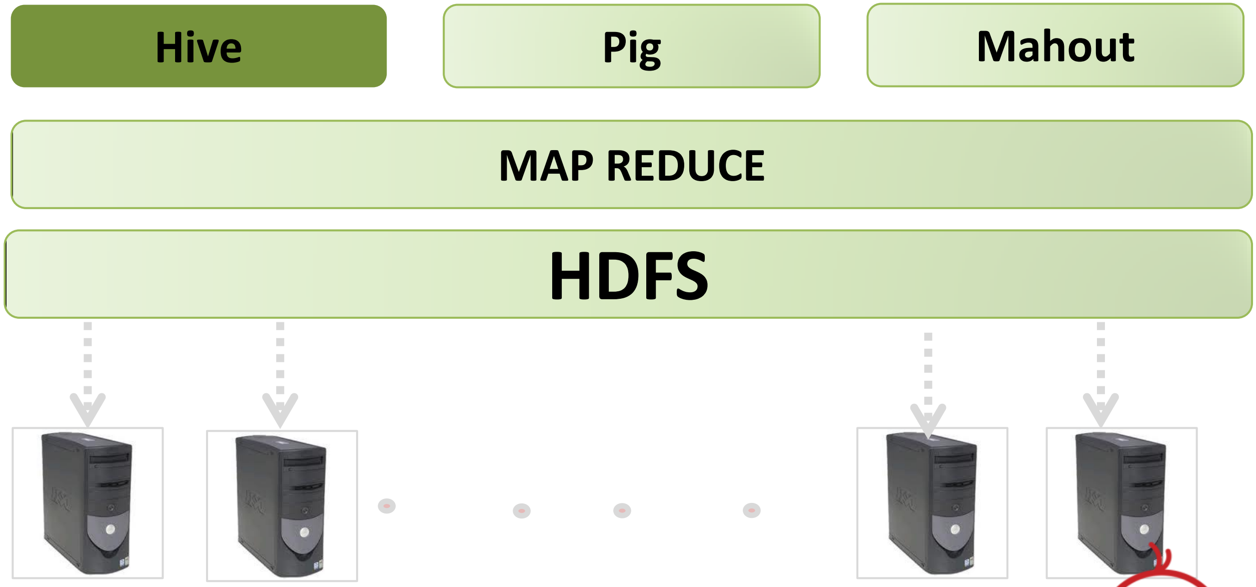
What About me???
I know only scripting and SQL.

- Busy Business Analyst

Hive

Originated from FaceBook

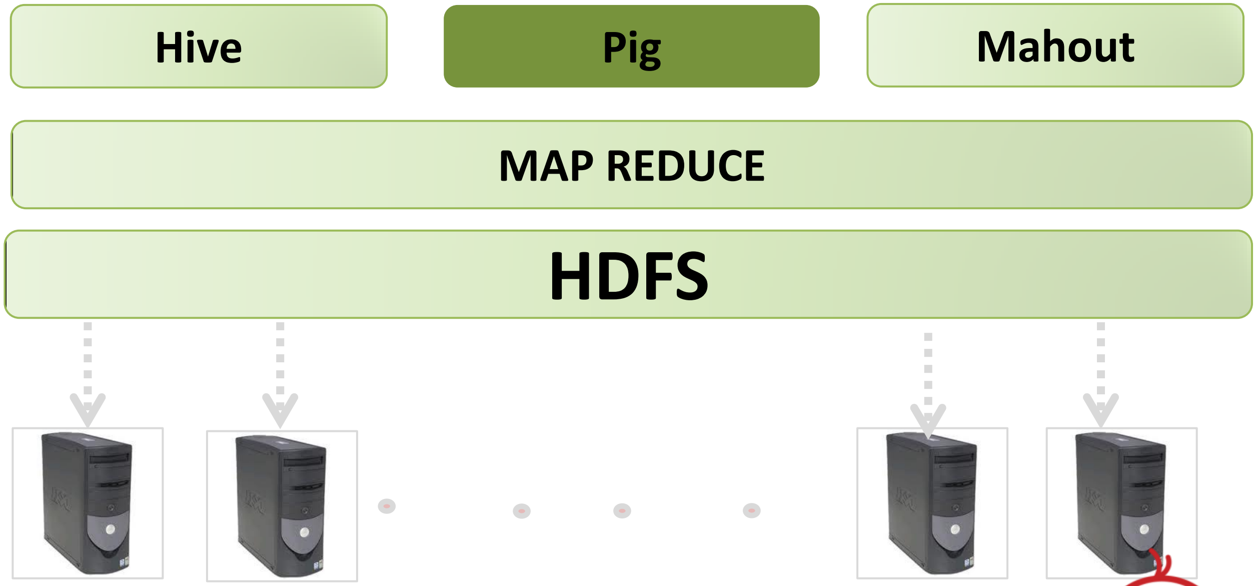
- Data Warehouse on Hadoop
- Structured Data Analysis
- Supports Subset of SQL-92 (HiveQL)
- SQL Queries into Map-Reduce



Pig

Originated from Yahoo

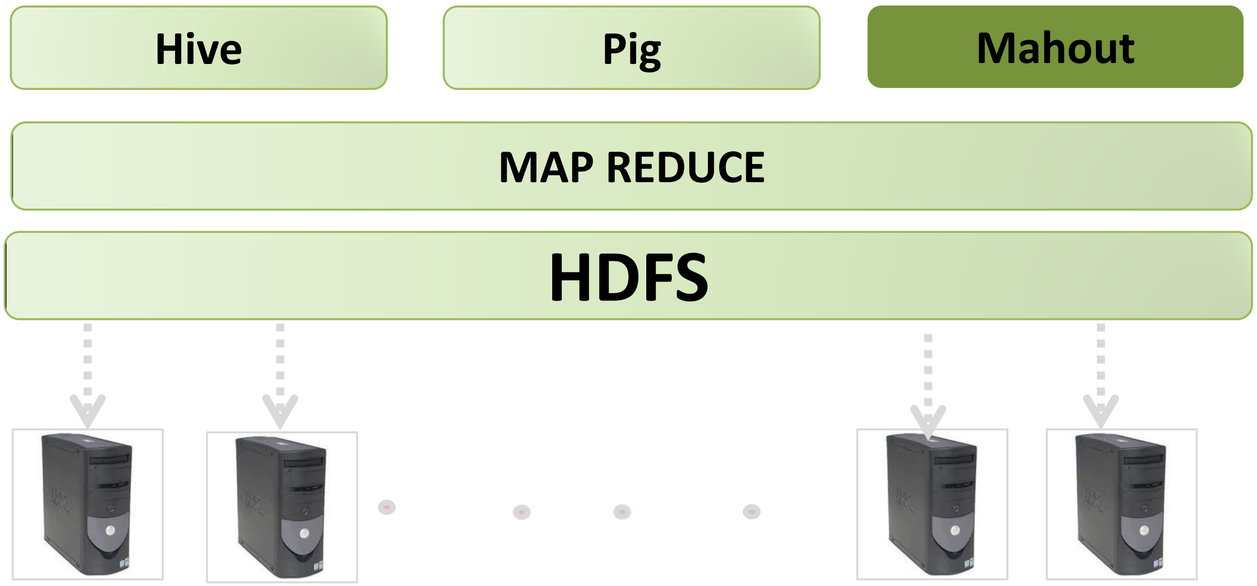
- Another Abstraction to MR like Hive
- DataFlow scripting Language (PigLatin)
- Meant for Data Factory usecases
- Can work on semistructured / strutured data



Mahout

Data Scientist !!!

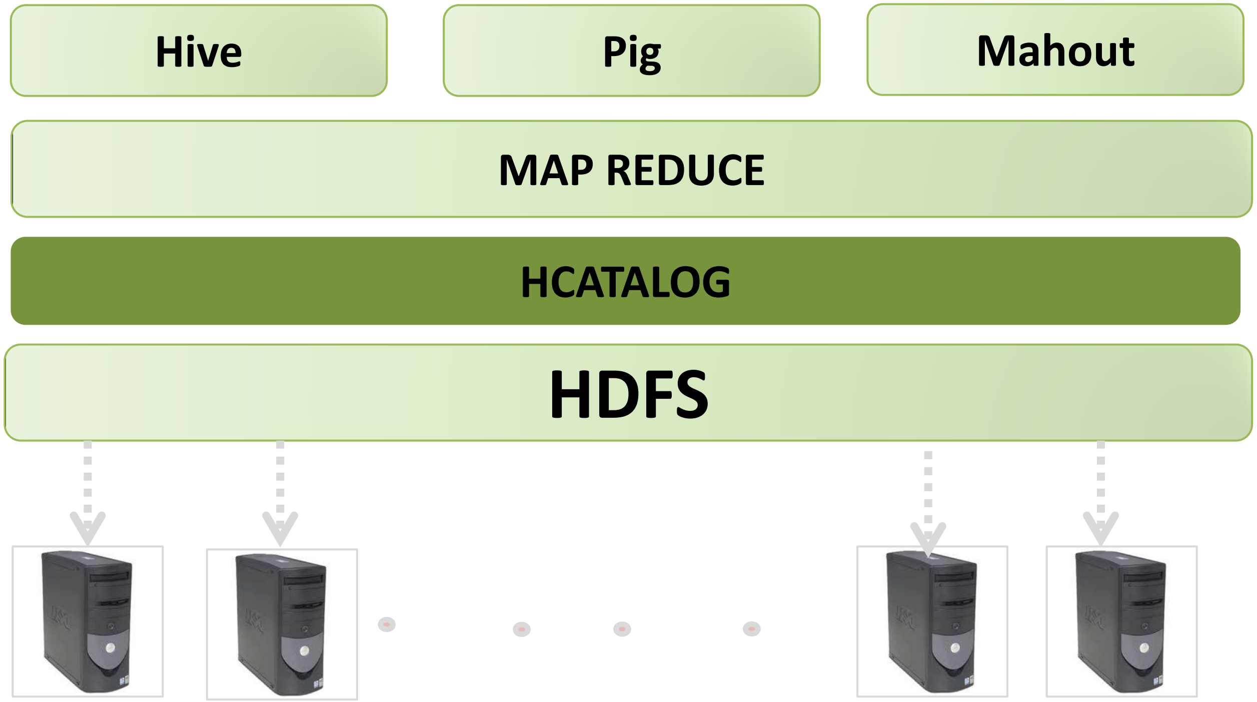
- A Set of Libraries
- Recommendations, Clustering, Classification, Collaborative filtering
- R , Python widely used on Hadoop



HCatalog

Table Management

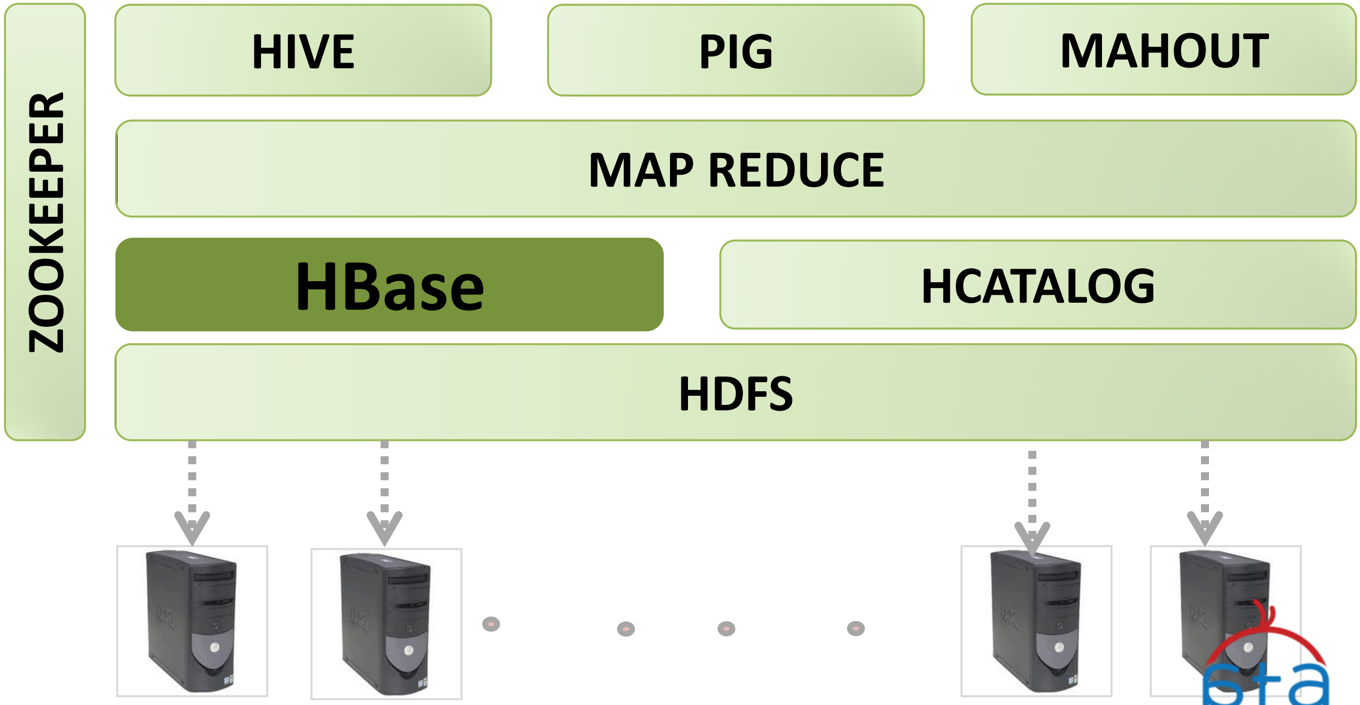
- Allows users to share the data and metadata across Hive , Pig and Others.
- Used by External tools such as Teradata Aster-H



HBase

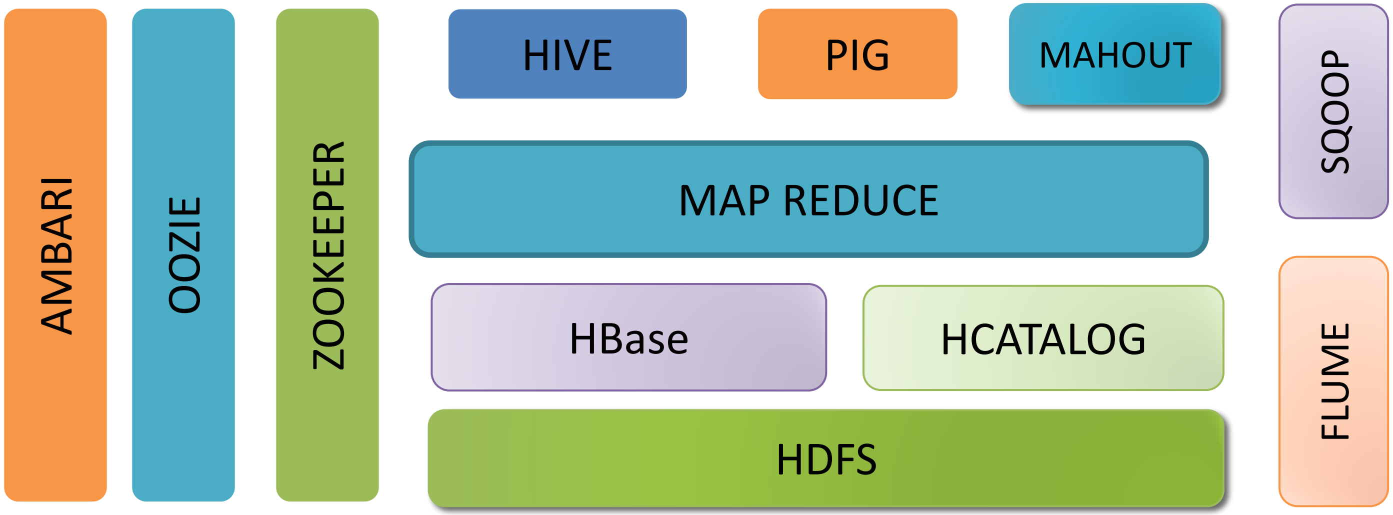
NoSQL

- = HDFS + Random read/writes
- Can be used for OLTP as well as OLAP applications
- Does not have secondary index by default



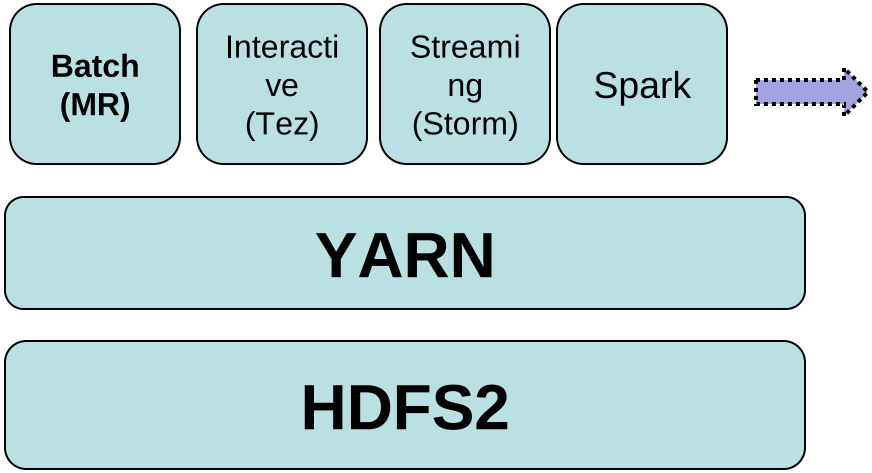
Big Picture – Till Yesterday

Its not Final !!!!

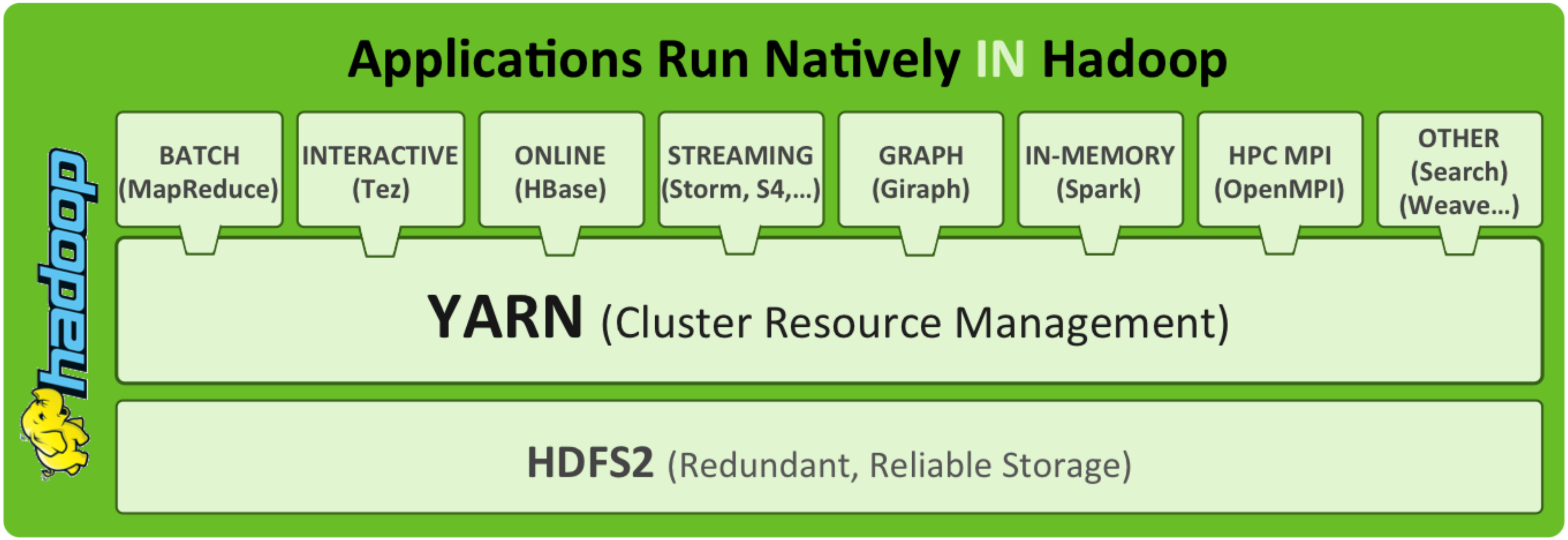


Update !!!

- Apache Hadoop YARN
  - Supports wide variety of applications such as batch, interactive, streaming .. etc
- High Availability for HDFS
  - Avoid SPOF
- HDFS Federation
- Support for Microsoft Windows.
- Snapshots for data.
- NFS-v3 Access.



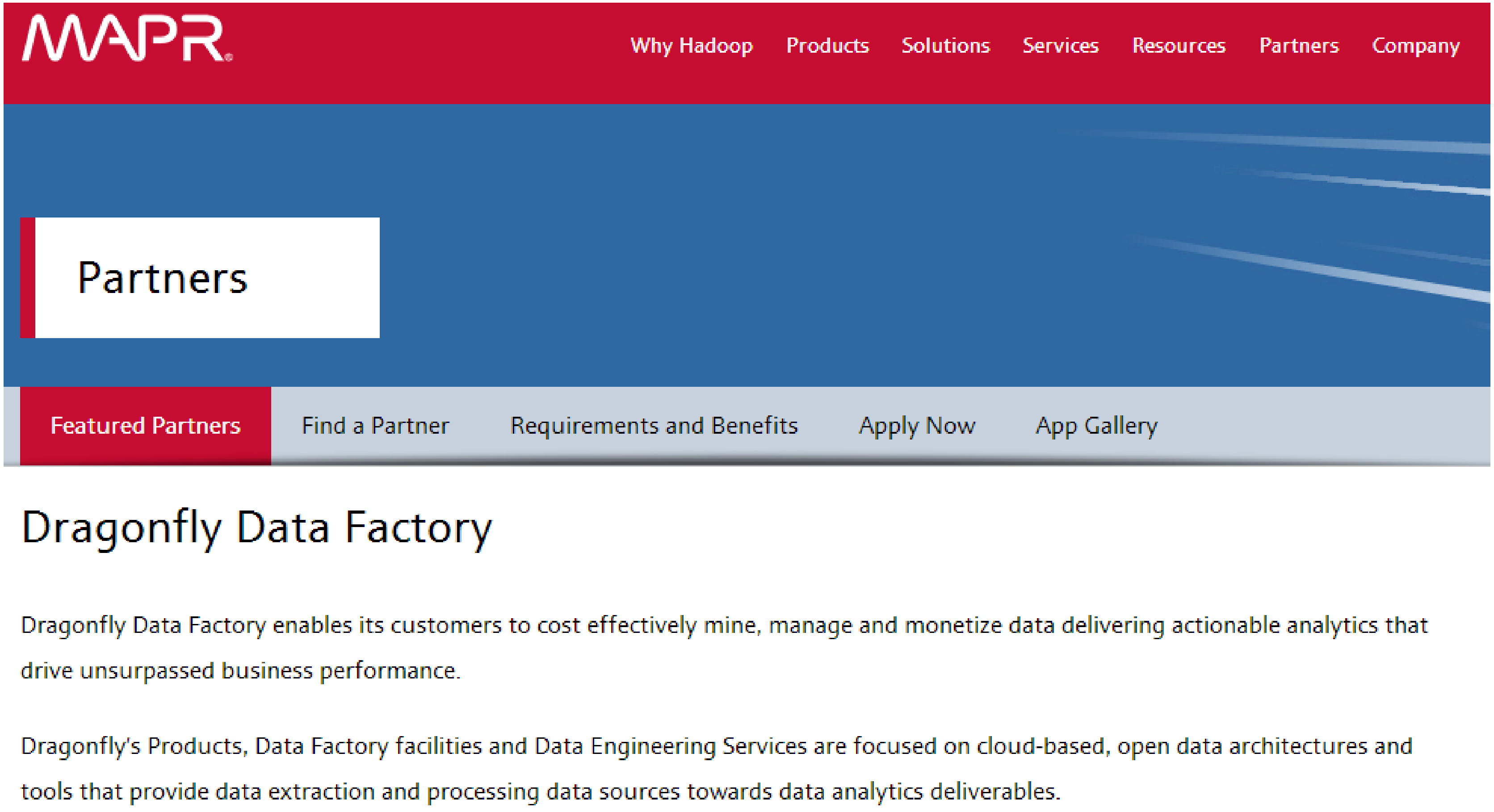
An updated Big Picture


Applications Run Natively IN Hadoop
hadoop
BATCH (MapReduce)
INTERACTIVE (Tez)
ONLINE (HBase)
STREAMING (Storm, S4,...)
GRAPH (Giraph)
IN-MEMORY (Spark)
HPC MPI (OpenMPI)
OTHER (Search) (Weave...)
YARN (Cluster Resource Management)
HDFS2 (Redundant, Reliable Storage)

Resource: www.hortonworks.com

Widely Used Hadoop Platforms

- Cloudera – CDH
- HortonWorks – HDP
- MapR – M3, M5, M7
- IBM – BigInsights
- DataStax
- EMC (GreenPlum) – PivotalHD
- Amazon Web Services – EMR
- WanDisco
- Intel - IDH



Partners

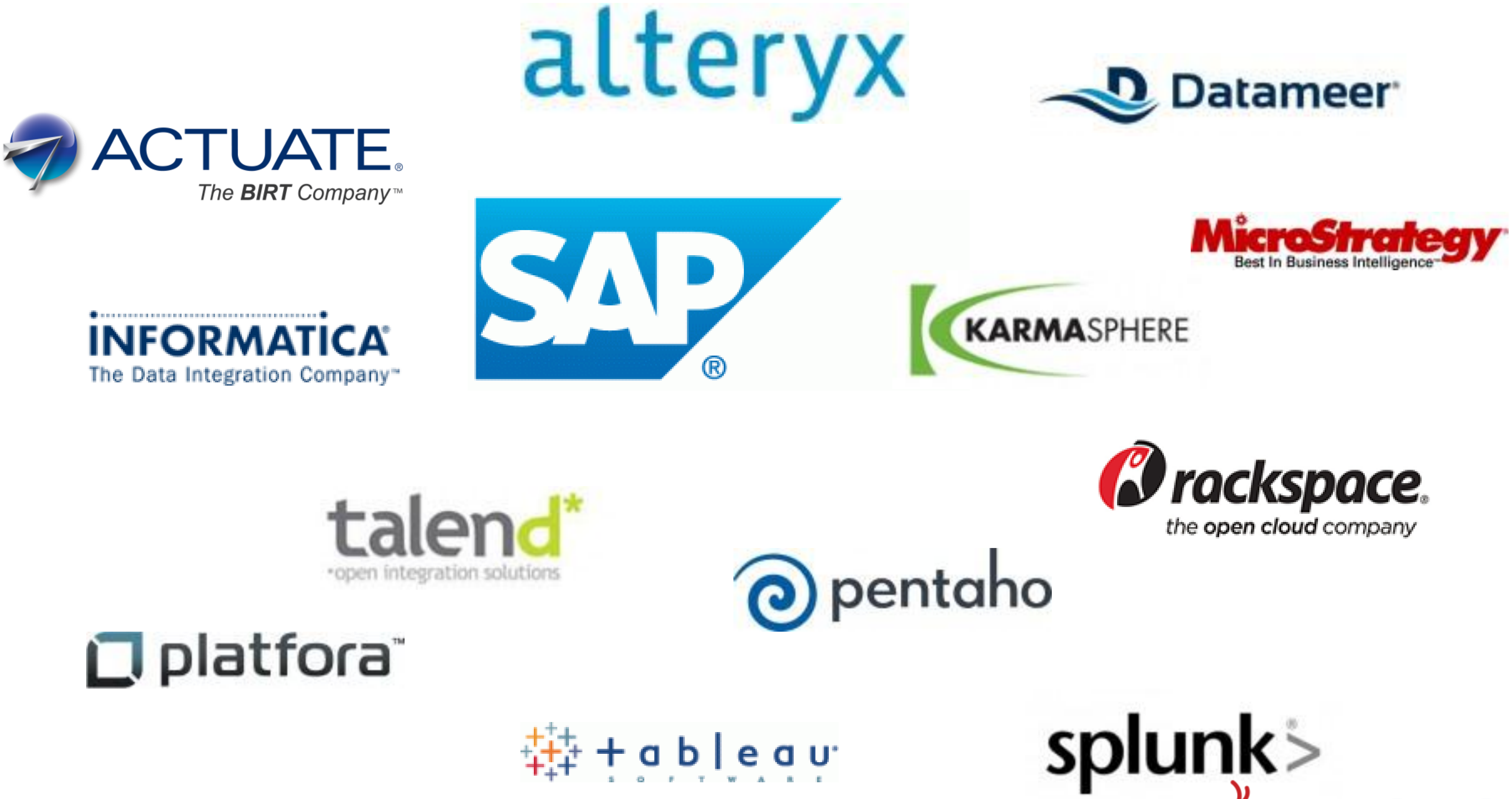
Dragonfly Data Factory

Dragonfly Data Factory enables its customers to cost effectively mine, manage and monetize data delivering actionable analytics that drive unsurpassed business performance.

Dragonfly's Products, Data Factory facilities and Data Engineering Services are focused on cloud-based, open data architectures and tools that provide data extraction and processing data sources towards data analytics deliverables.

Hadoop Connectors from Existing Products

Sample List



Thanks Man!!

But Hadoop == BigData??

HADOOP != BIG DATA

A de-facto standard for solving the most of the problems of BigData

Other Big Data Technologies

- NoSQL – Cassandra, MongoDB, CouchDB, DynamoDB, Riak, MarkLogic
- Log Aggregators - Kafka, Scribe,LogStash, GrayLog2
- Search Analytics – Lucene (ElasticSearch, Solr)
- Analytics - Rhadoop, RHIPE
- Stream Processing - STORM, Samza , S4, Muppet
- InMemory DataGrid - Memcache

Thank You

DATA DOTZ