BigData

An Introduction

Agenda

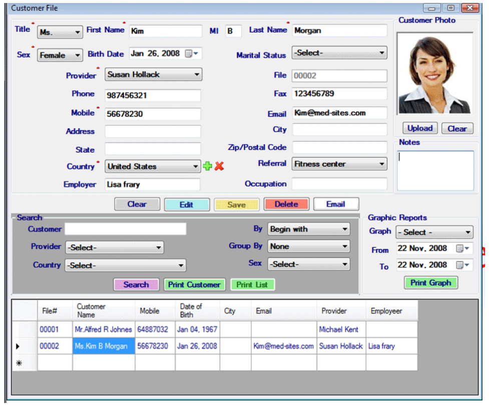
- The Data Voyage
- Understanding BigData
- BigData Enablers
- BigData use cases
- BigData Market

TLS

The Data Voyage

Understanding Data

What is data? A Customer’s record, Subscriber’s telephone # and address, a retail ticket, an insurance claim, an inventory list, a sales order, a flight ticket…etc.



Customer File

Title: Ms. First Name: Kim MI: B Last Name: Morgan

Sex: Female Birth Date: Jan 26, 2008 Marital Status: -Select-

Provider: Susan Hollack File: 00002

Phone: 987456321 Fax: 123456789

Mobile: 56678230 Email: Kim@med-sites.com

Address: City:

State: Zip/Postal Code:

Country: United States Referral: Fitness center

Employer: Lisa frary Occupation:

Customer Photo

Notes:

Search

Customer: By: Begin with

Provider: -Select- Group By: None

Country: -Select- Sex: -Select-

Graphic Reports

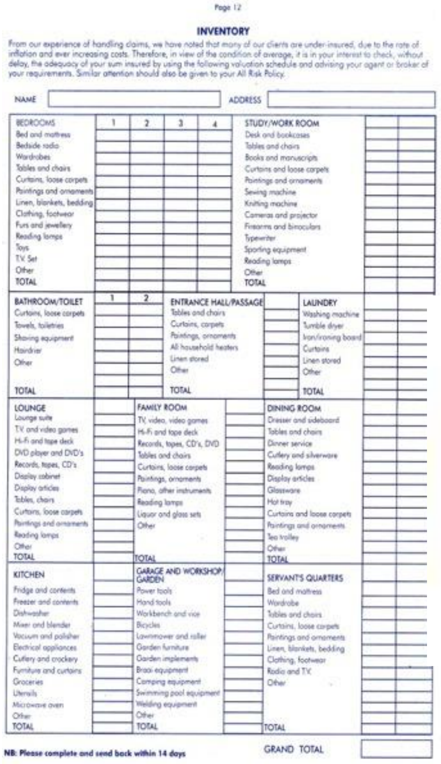
Graph: - Select -

From: 22 Nov. 2008

To: 22 Nov. 2008

	File#	Customer Name	Mobile	Date of Birth	City	Email	Provider	Employeer
	00001	Mr.Alfred R Johnes	64887032	Jan 04, 1967			Michael Kent	
▸	00002	Ms.Kim B Morgan	56678230	Jan 26, 2008		Kim@med-sites.com	Susan Hollack	Lisa frary
*								

Upload
Clear
Clear
Edit
Save
Delete
Email
Search
Print Customer
Print List
Print Graph



Page 12

INVENTORY

From our experience of handling claims, we have noted that many of our clients are under-insured, due to the rate of inflation and ever increasing costs. Therefore, in view of the condition of average, it is in your interest to check, without delay, the adequacy of your sum insured by using the following valuation schedule and advising your agent or broker of your requirements. Similar attention should also be given to your All Risk Policy.

NAME ADDRESS

BEDROOMS 1 2 3 4
Bed and mattress
Bedside radio
Wardrobes
Tables and chairs
Curtains, loose carpets
Paintings and ornaments
Linen, blankets, bedding
Clothing, footwear
Furs and jewellery
Reading lamps
Toys
T.V. Set
Other
TOTAL

STUDY/WORK ROOM
Desk and bookcases
Tables and chairs
Books and manuscripts
Curtains and loose carpets
Paintings and ornaments
Sewing machine
Knitting machine
Cameras and projector
Firearms and binoculars
Typewriter
Sporting equipment
Reading lamps
Other
TOTAL

BATHROOM/TOILET 1 2
Curtains, loose carpets
Towels, toiletries
Shaving equipment
Hairdrier
Other
TOTAL

ENTRANCE HALL/PASSAGE
Tables and chairs
Curtains, carpets
Paintings, ornaments
All household heaters
Linen stored
Other
TOTAL

LAUNDRY
Washing machine
Tumble dryer
Iron/ironing board
Curtains
Linen stored
Other
TOTAL

LOUNGE
Lounge suite
T.V. and video games
Hi-Fi and tape deck
DVD player and DVD's
Records, tapes, CD's
Display cabinet
Display articles
Tables, chairs
Curtains, loose carpets
Paintings and ornaments
Reading lamps
Other
TOTAL

FAMILY ROOM
TV, video, video games
Hi-Fi and tape deck
Records, tapes, CD's, DVD
Tables and chairs
Curtains, loose carpets
Paintings, ornaments
Piano, other instruments
Reading lamps
Liquor and glass sets
Other
TOTAL

DINING ROOM
Dresser and sideboard
Tables and chairs
Dinner service
Cutlery and silverware
Reading lamps
Display articles
Glassware
Hot tray
Curtains and loose carpets
Paintings and ornaments
Tea trolley
Other
TOTAL

KITCHEN
Fridge and contents
Freezer and contents
Dishwasher
Mixer and blender
Vacuum and polisher
Electrical appliances
Cutlery and crockery
Furniture and curtains
Groceries
Utensils
Microwave oven
Other
TOTAL

GARAGE AND WORKSHOP/GARDEN
Power tools
Hand tools
Workbench and vice
Bicycles
Lawnmower and roller
Garden furniture
Garden implements
Braai equipment
Camping equipment
Swimming pool equipment
Welding equipment
Other
TOTAL

SERVANT'S QUARTERS
Bed and mattress
Wardrobe
Tables and chairs
Curtains, loose carpets
Paintings and ornaments
Linen, blankets, bedding
Clothing, footwear
Radio and T.V.
Other
TOTAL

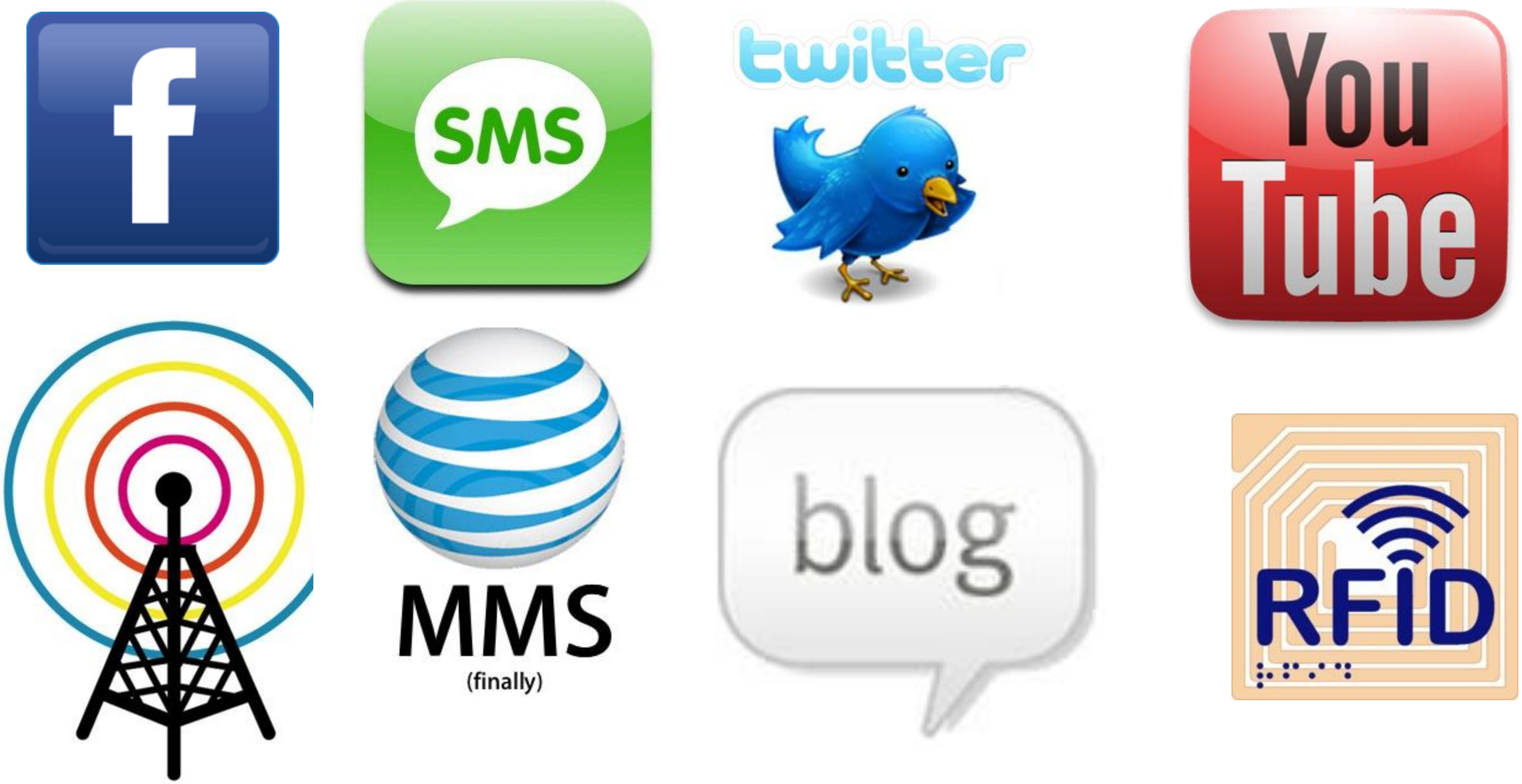
NB: Please complete and send back within 14 days

GRAND TOTAL

This is how we perceive data today. But data is changing its form, even right now as we speak, data is changing.

Data Amorphism

The manifestation of data has changed dramatically over the last few years. Data today is quite different from what data used to be a few years ago!



The Data Challenge

- This new data is different from the regular data in a couple of respects, not only is it often unstructured, it has high volumes and in many cases changes rapidly
- As Gartner defines it, this new data has
- High Volume
- High Velocity
- High Variety
- High Complexity

The Data Challenge… contd.

The industry has tried to tackle this “Data Challenge” from different technology and process perspectives. Some methods to improve the handling of data and others to directly improve the subsequent usage (Analytics) of it.

- Grid Computing (Volume)
- Cloud Computing (Velocity & Volume)
- Using Open Source Software (overall cost)
- Virtualization (Time to Test/Deploy, Processing Capacity)
- Innovations in BI – Realtime & Operational (Velocity)
- Dw Appliances (Volume)
- Industry specific models and solutions (Time to Deploy, Complexity)

The Paradigm Shift

- Clouds are slow, Grids are expensive, Open Source less robust, Virtual Servers tend to slow things down, Dw Appliances bring in a lot of hardware…

Hadoop is an Open Source data storage and processing Architecture.

TLS

The Paradigm Shift…contd.

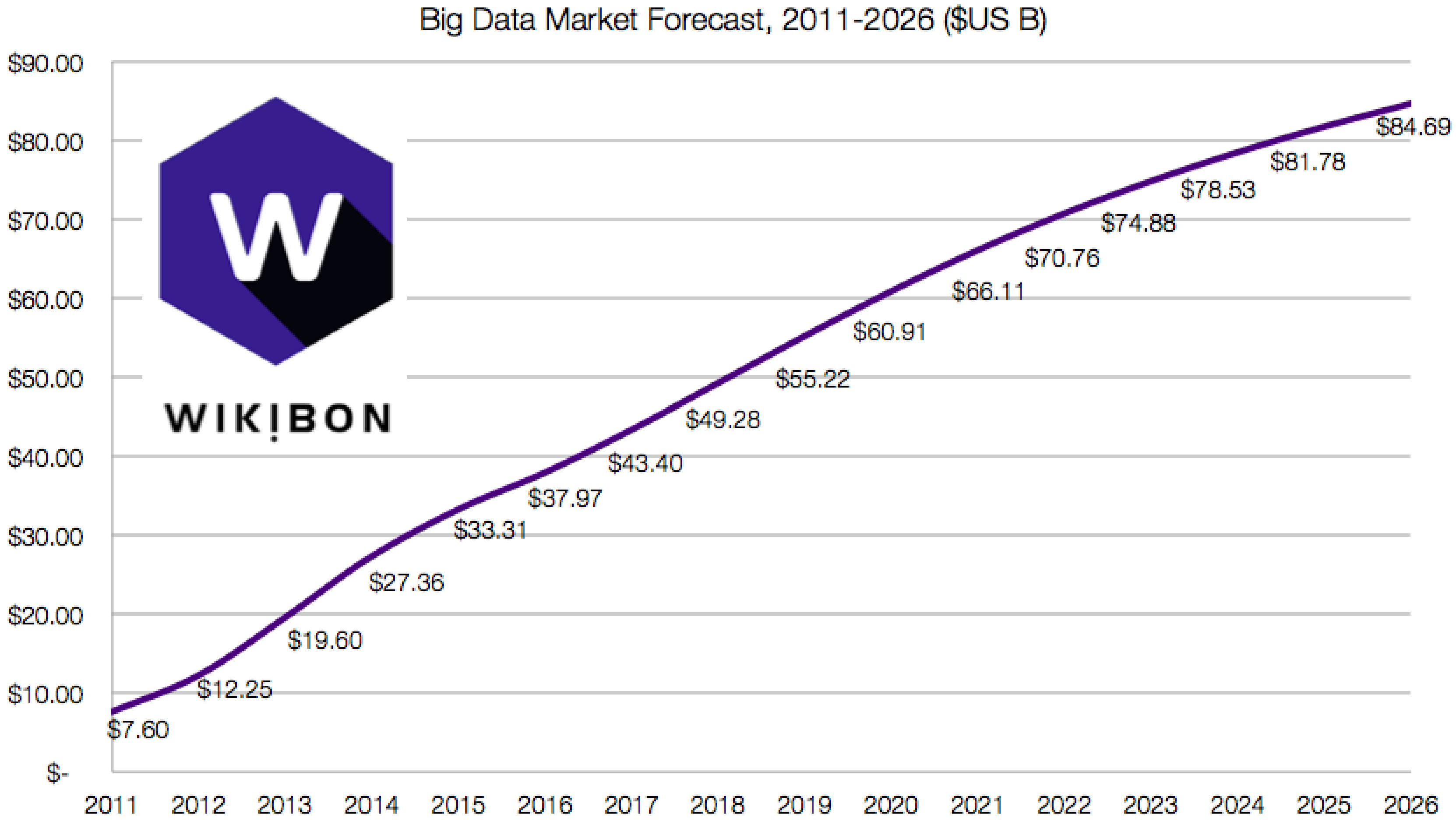
By the early part of the last decade, Internet based Cos. Like Google, Yahoo, Linkedin, Amazon and others adopted Bigdata and invented programming models, frameworks to suit their needs and fill the knowledge gap.

With Hadoop & its associated technologies they could

- Handle terabytes & petabytes of data
- which was unstructured – text, audio, video & various formats
- on commodity hardware
- assuring fail-safe (read and write) operations
- with high performance
- and economy (remember its Open Source!)

More on Hadoop in the coming slides.

Some facts - Wikibon

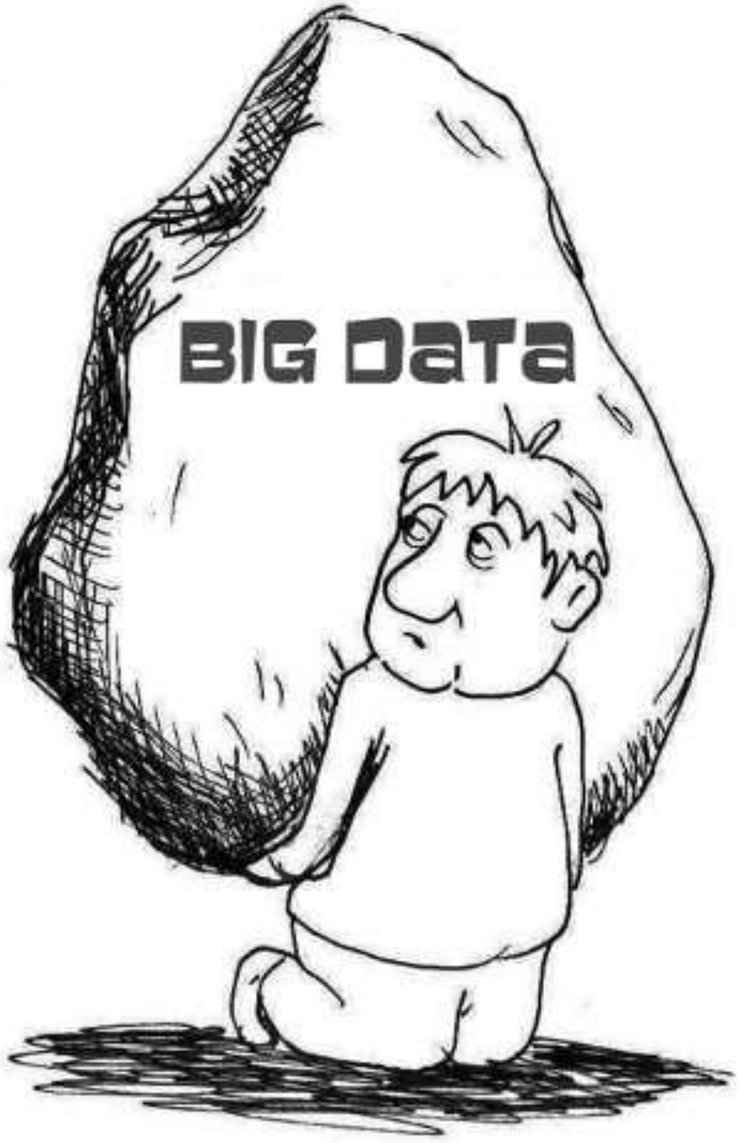


TLS

Some facts - eWeek

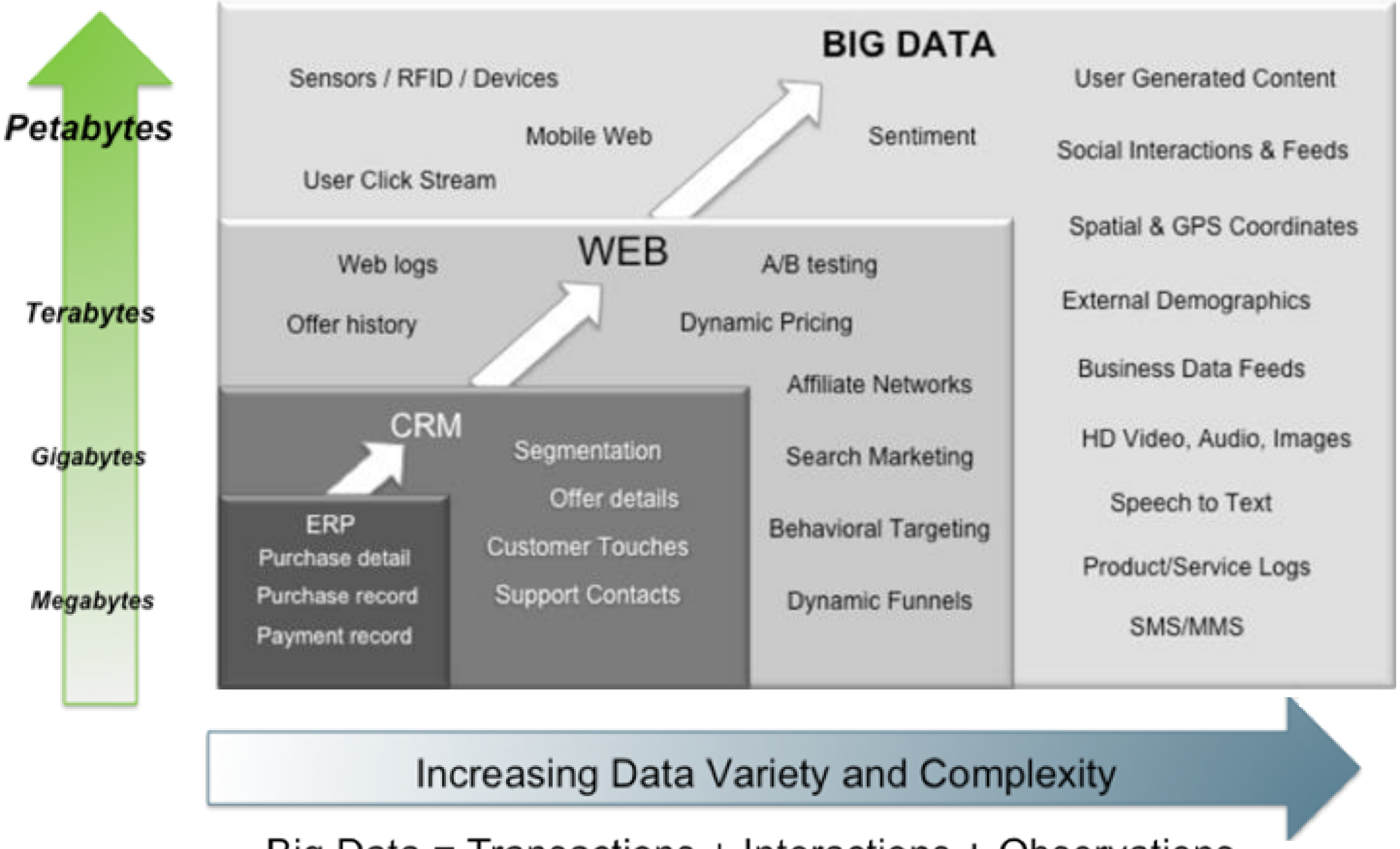
- Allied Market Research forecast that the global Hadoop market value will reach **$50.2 billion by 2020**
- Hadoop Will Be Used for Over 10 Percent of Data Processing and Storage
- Hadoop Will Lead in Infrastructure Spending
- Hadoop Will Not Be the Greatest Expense of the IT Budget
- Hadoop Will Democratize Data
- Hadoop Will Accelerate Big Data Adoption
- Hadoop Will Power the Startup Economy
- Hadoop Will Be Used for Critical Day-to-Day
- Hadoop Will Advance the Internet of Things
- Hadoop Will Be Used for Processing and Storing Highly Sensitive Data
- Hadoop Will Power the Connected World

http://www.eweek.com/database/slideshows/hadoop-2020-the-future-of-big-data-in-the-enterprise.html#sthash.FJ5FW6Kt.dpuf



Understanding Bigdata

BigData

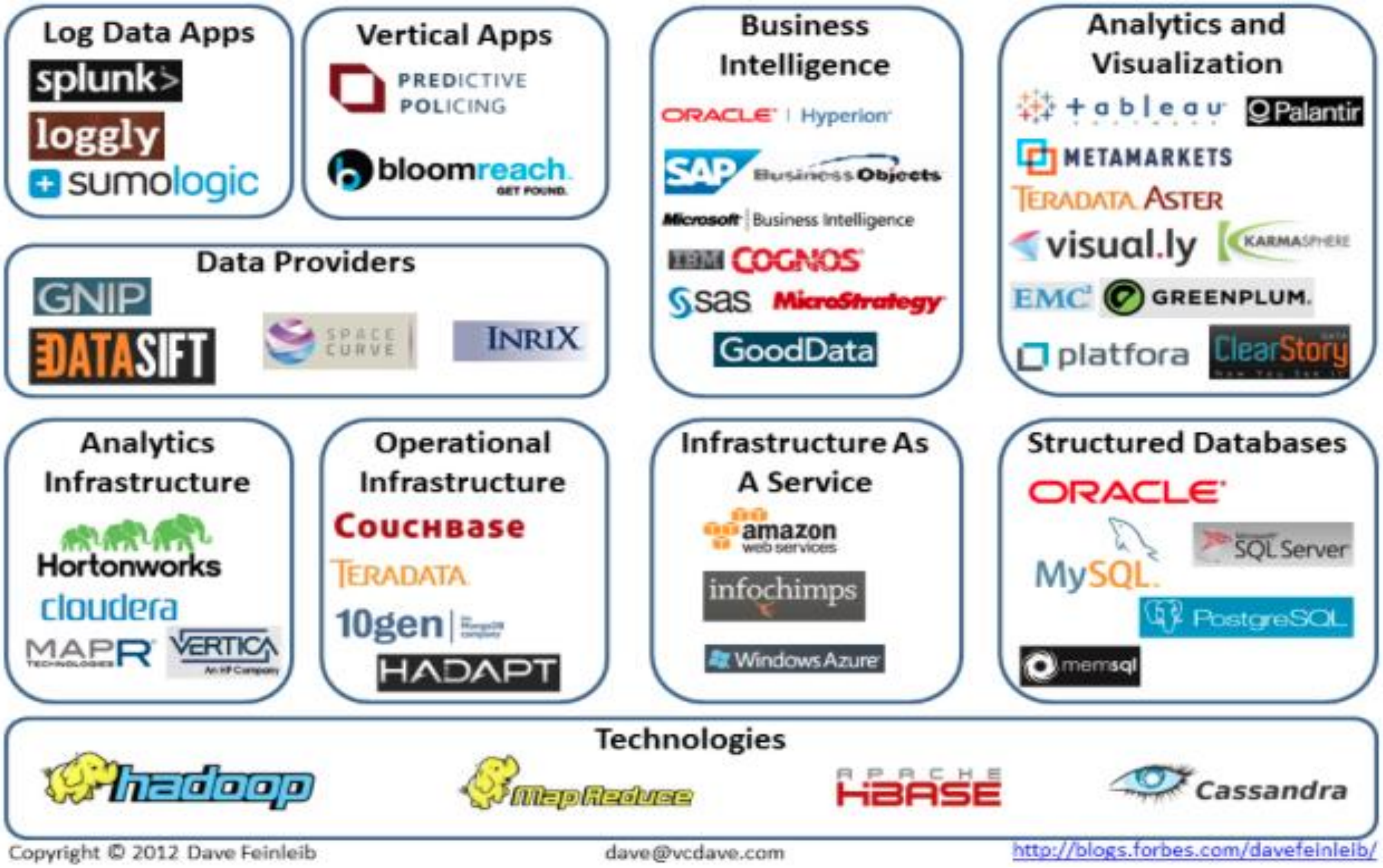


Big Data = Transactions + Interactions + Observations

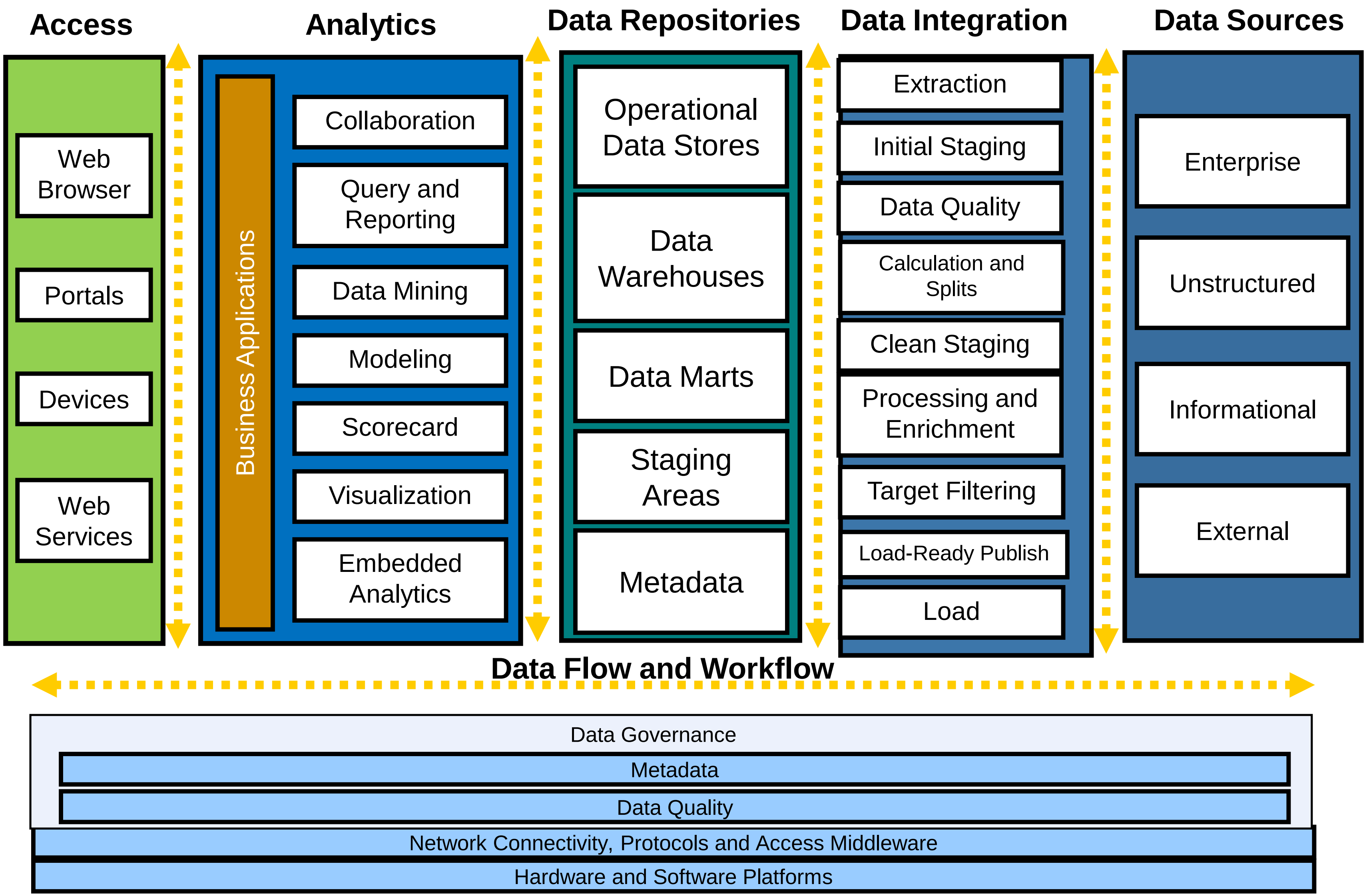
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

TLS

Bigdata Technology Landscape



Conventional BI Landscape



Access

- Web Browser
- Portals
- Devices
- Web Services

Analytics

- Business Applications
- Collaboration
- Query and Reporting
- Data Mining
- Modeling
- Scorecard
- Visualization
- Embedded Analytics

Data Repositories

- Operational Data Stores
- Data Warehouses
- Data Marts
- Staging Areas
- Metadata

Data Integration

- Extraction
- Initial Staging
- Data Quality
- Calculation and Splits
- Clean Staging
- Processing and Enrichment
- Target Filtering
- Load-Ready Publish
- Load

Data Sources

- Enterprise
- Unstructured
- Informational
- External

Data Flow and Workflow

Data Governance

Metadata

Data Quality

Network Connectivity, Protocols and Access Middleware

Hardware and Software Platforms

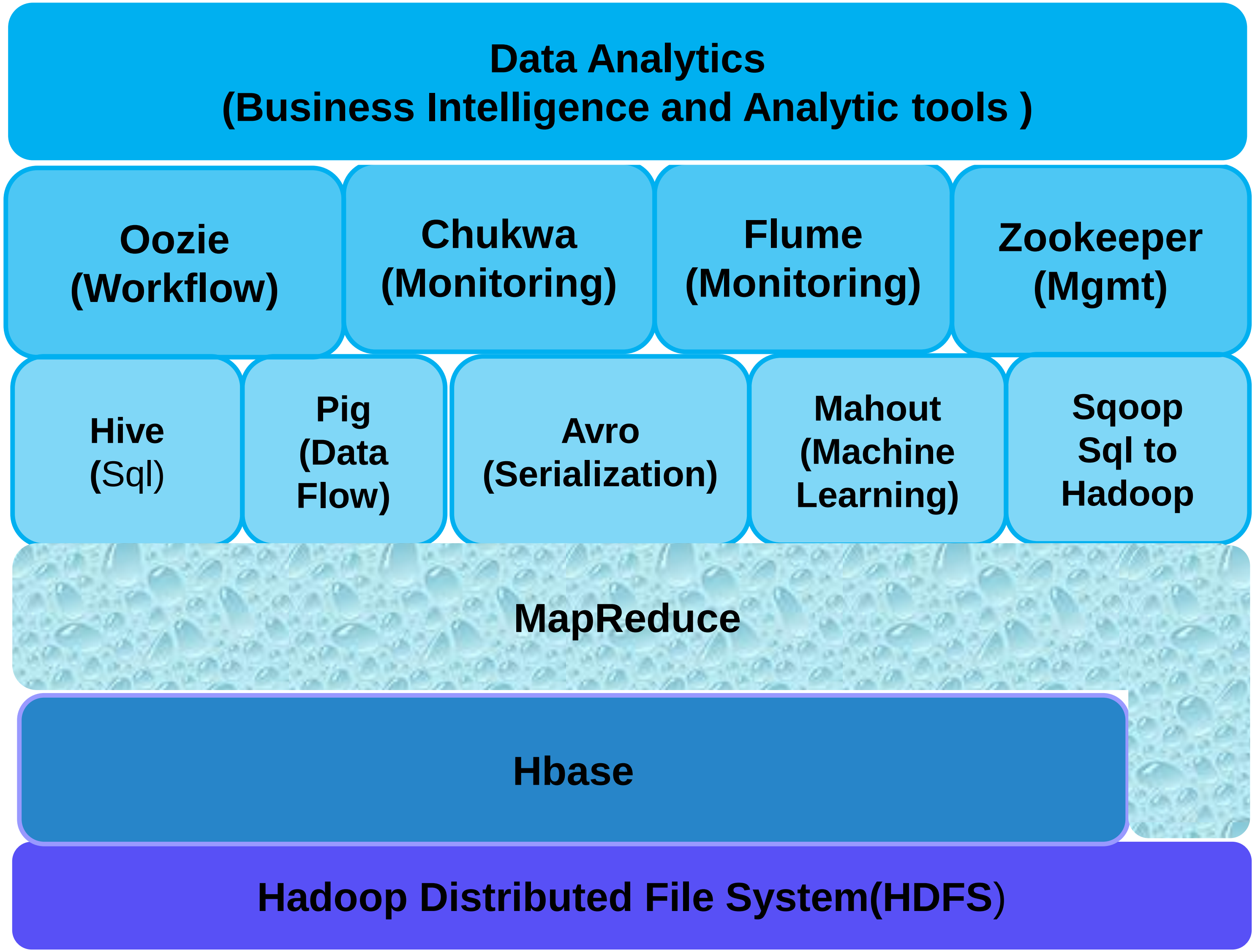
What happens to the old data order?

Is this the end of the road for Conventional BI?

Hadoop Clusters–Storage blocks of Bigdata

Cluster of machines running Hadoop at Yahoo (Source: Yahoo)

The Hadoop Ecosystem



Data Analytics
(Business Intelligence and Analytic tools)

Oozie
(Workflow)

Chukwa
(Monitoring)

Flume
(Monitoring)

Zookeeper
(Mgmt)

Hive
(Sql)

Pig
(Data
Flow)

Avro
(Serialization)

Mahout
(Machine
Learning)

Sqoop
Sql to
Hadoop

MapReduce

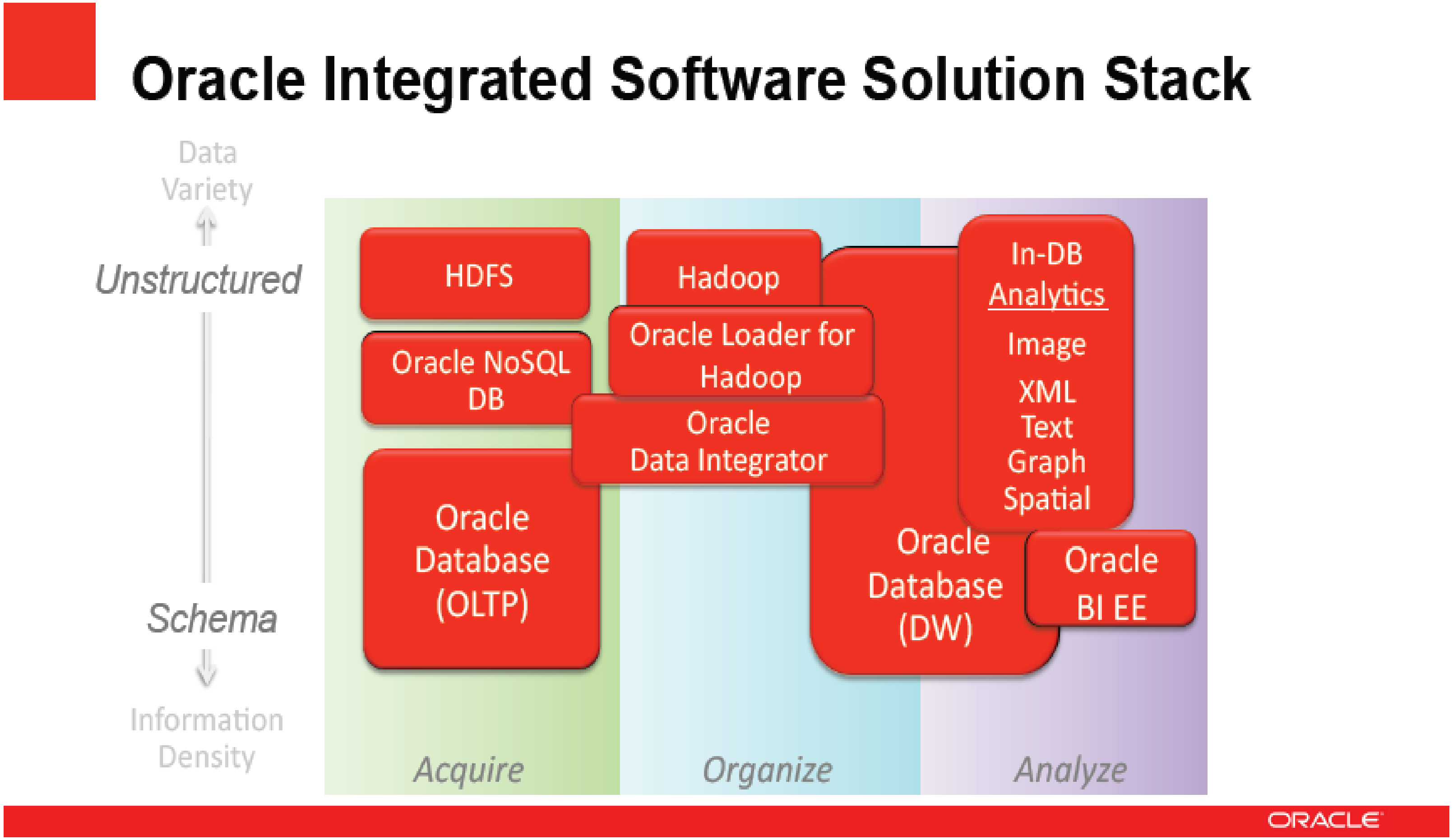
Hbase

Hadoop Distributed File System(HDFS)

TLS

Bigdata Enablers

Oracle Bigdata Reference Landscape



Oracle Integrated Software Solution Stack

Data Variety

Unstructured

Schema

Information Density

HDFS

Oracle NoSQL DB

Oracle Database (OLTP)

Hadoop

Oracle Loader for Hadoop

Oracle Data Integrator

Oracle Database (DW)

In-DB Analytics

Image

XML

Text

Graph

Spatial

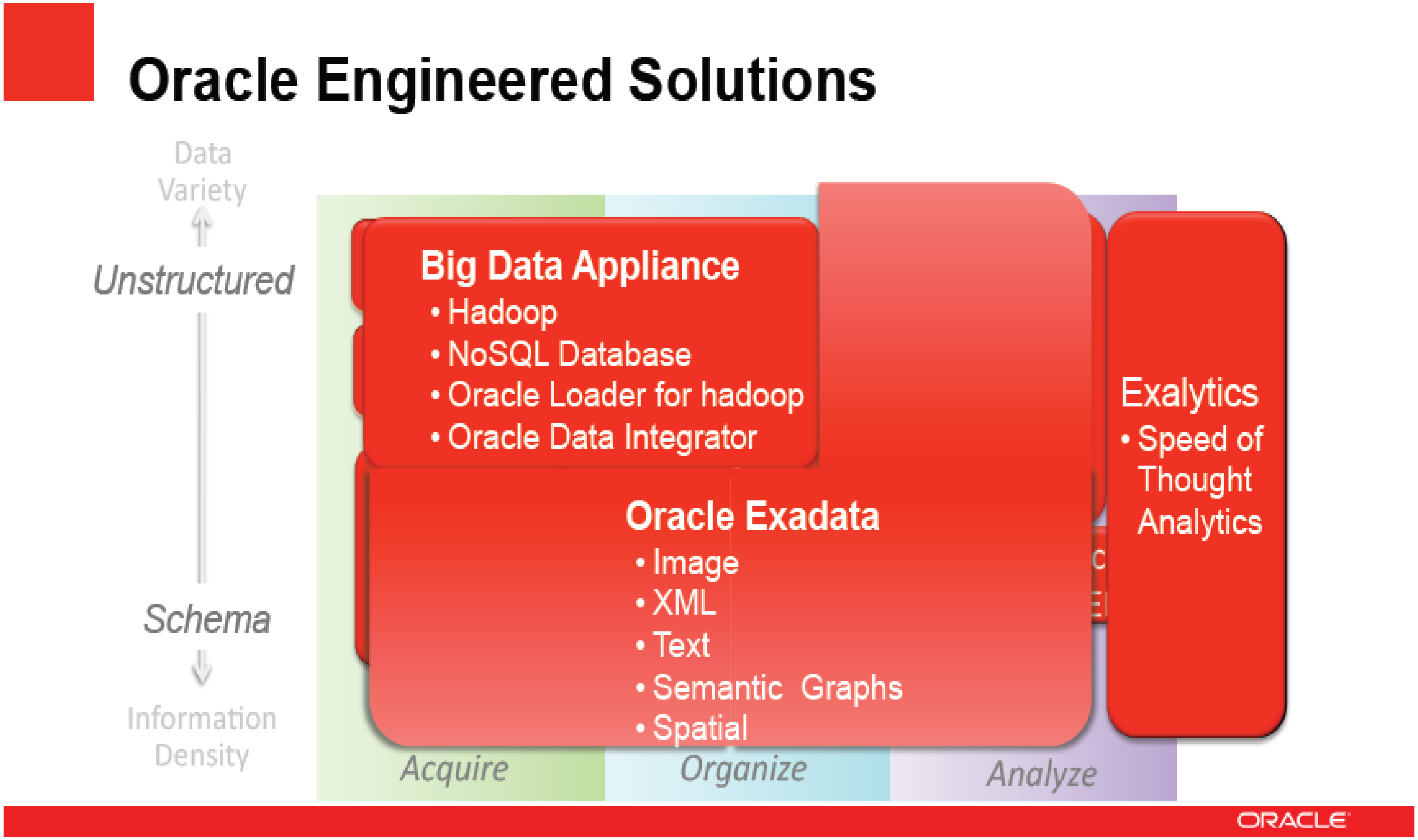
Oracle BI EE

Acquire

Organize

Analyze

Oracle Bigdata Reference Landscape



Oracle Engineered Solutions

Data Variety

Unstructured

Schema

Information Density

Big Data Appliance

- Hadoop
- NoSQL Database
- Oracle Loader for hadoop
- Oracle Data Integrator

Oracle Exadata

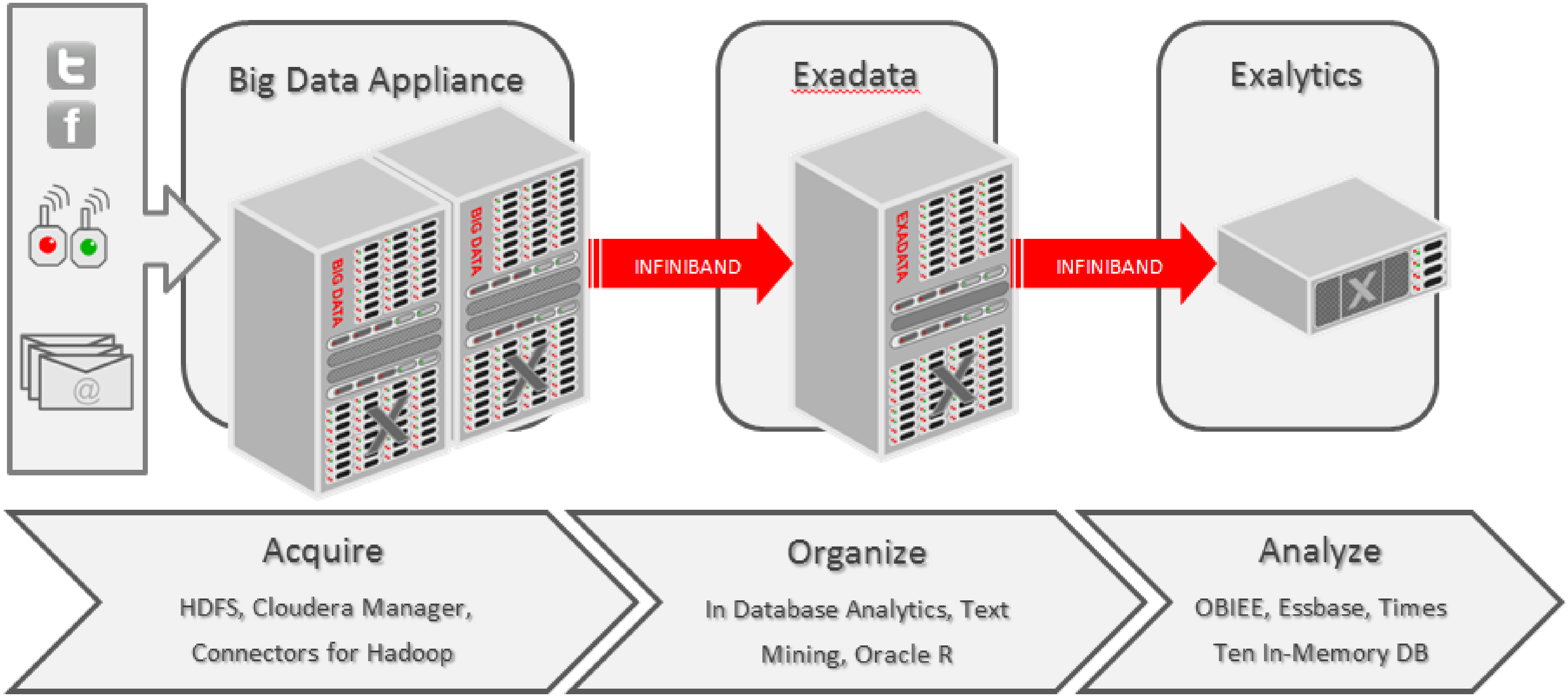
- Image
- XML
- Text
- Semantic Graphs
- Spatial

Exalytics

- Speed of Thought Analytics

Oracle Bigdata Reference Landscape

Another perspective of Oracle BigData Solution



Acquire

HDFS, Cloudera Manager, Connectors for Hadoop

Organize

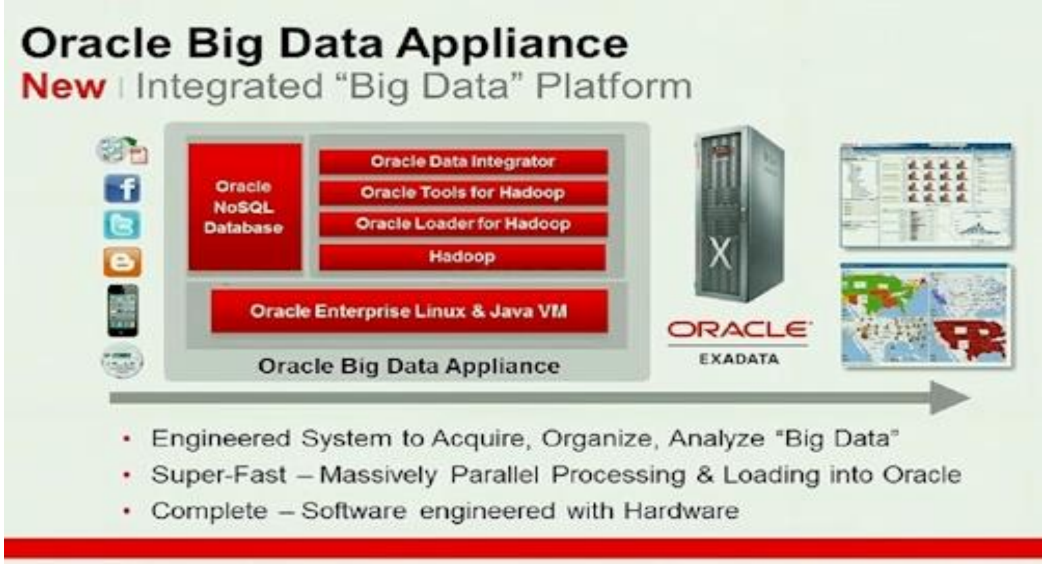
In Database Analytics, Text Mining, Oracle R

Analyze

OBIEE, Essbase, Times Ten In-Memory DB

Oracle Bigdata Appliance Architecture

- The Oracle Bigdata Appliance software includes
- Full distribution of Cloudera’s version of Hadoop including Apache Hadoop (CDH4)
- Oracle Bigdata Appliance Plug-In for Enterprise Manager
- Cloudera Manager to administer all aspects of Cloudera CDH
- Oracle distribution of the statistical package R
- Oracle NoSQL Database Community Edition2
- Oracle Enterprise Linux operating system and Oracle Java VM



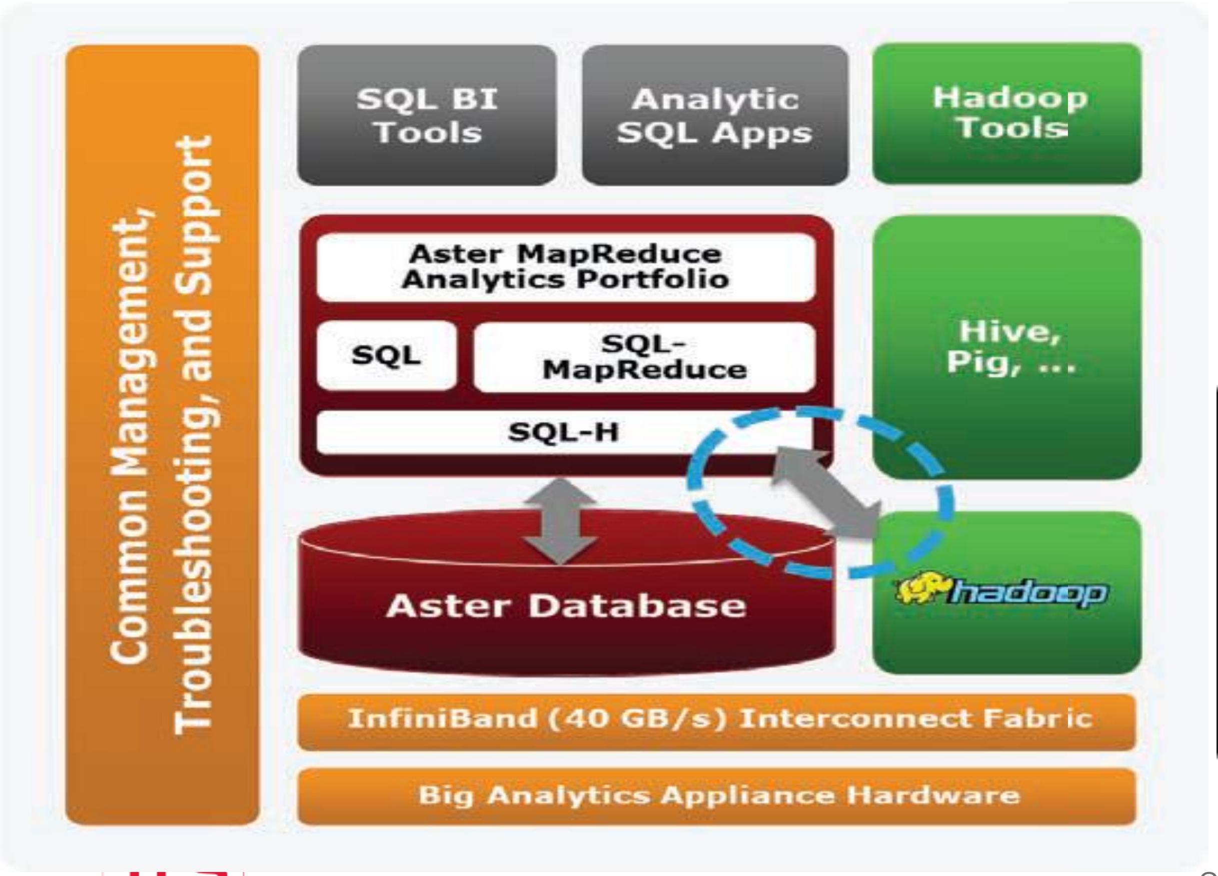
Oracle Big Data Appliance

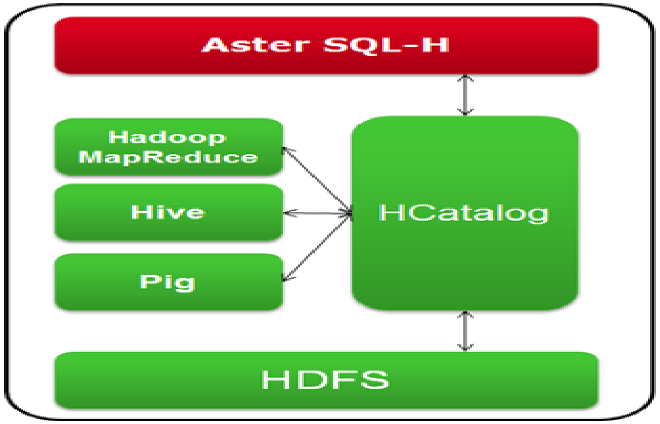
New | Integrated “Big Data” Platform

- Engineered System to Acquire, Organize, Analyze “Big Data”
- Super-Fast – Massively Parallel Processing & Loading into Oracle
- Complete – Software engineered with Hardware

Teradata Aster Big Analytics Appliance

The Teradata Aster Big Analytics appliance offers a Hybrid architecture that includes Aster Database, Aster SQL-mapreduce and Apache Hadoop

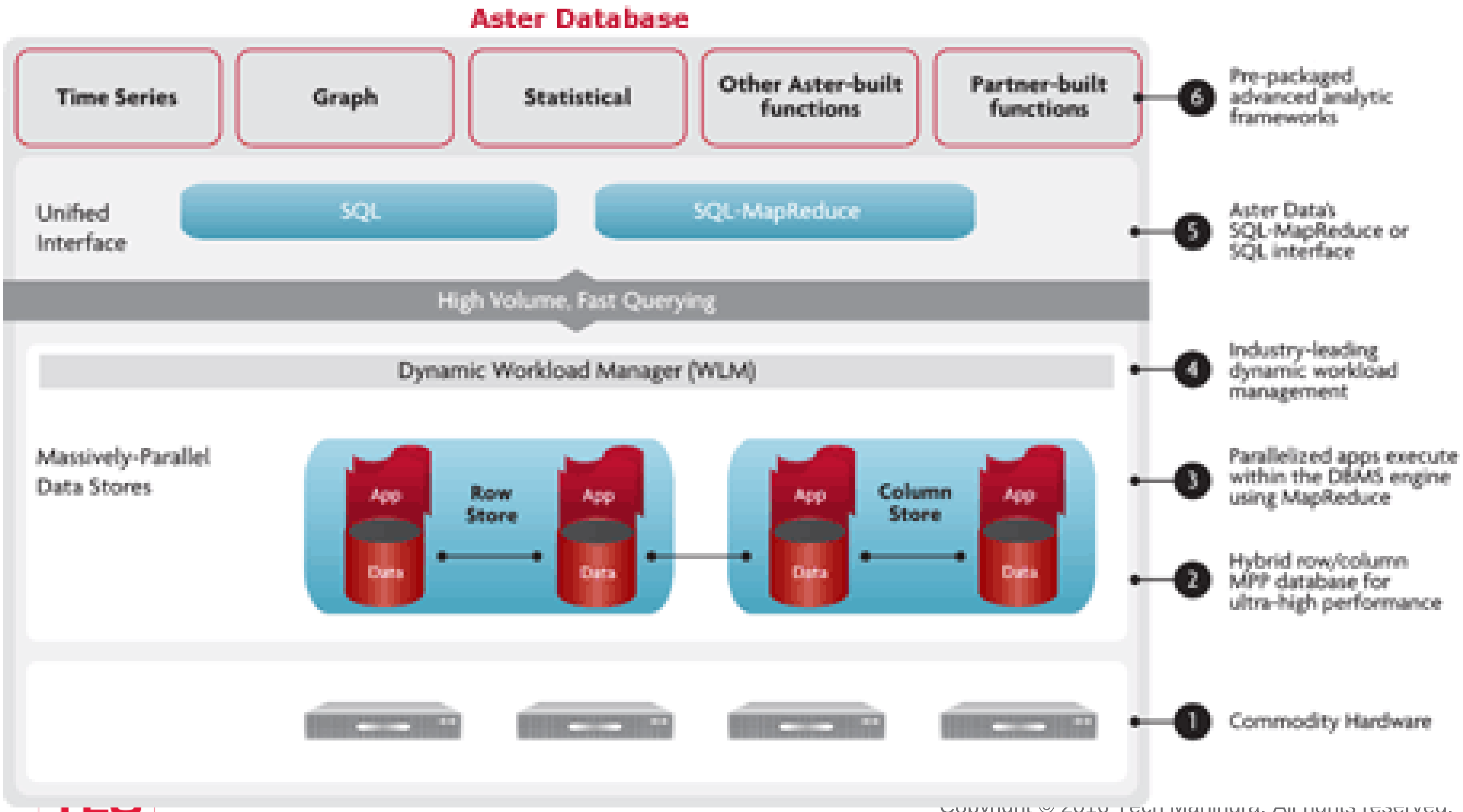

Common Management, Troubleshooting, and Support
SQL BI Tools
Analytic SQL Apps
Hadoop Tools
Aster MapReduce Analytics Portfolio
SQL
SQL-MapReduce
SQL-H
Hive, Pig, ...
Aster Database
hadoop
InfiniBand (40 GB/s) Interconnect Fabric
Big Analytics Appliance Hardware



SQL-H interfaces with the Apache HCatalog to provide a mechanism for users to directly access the data in Hadoop from Aster Database

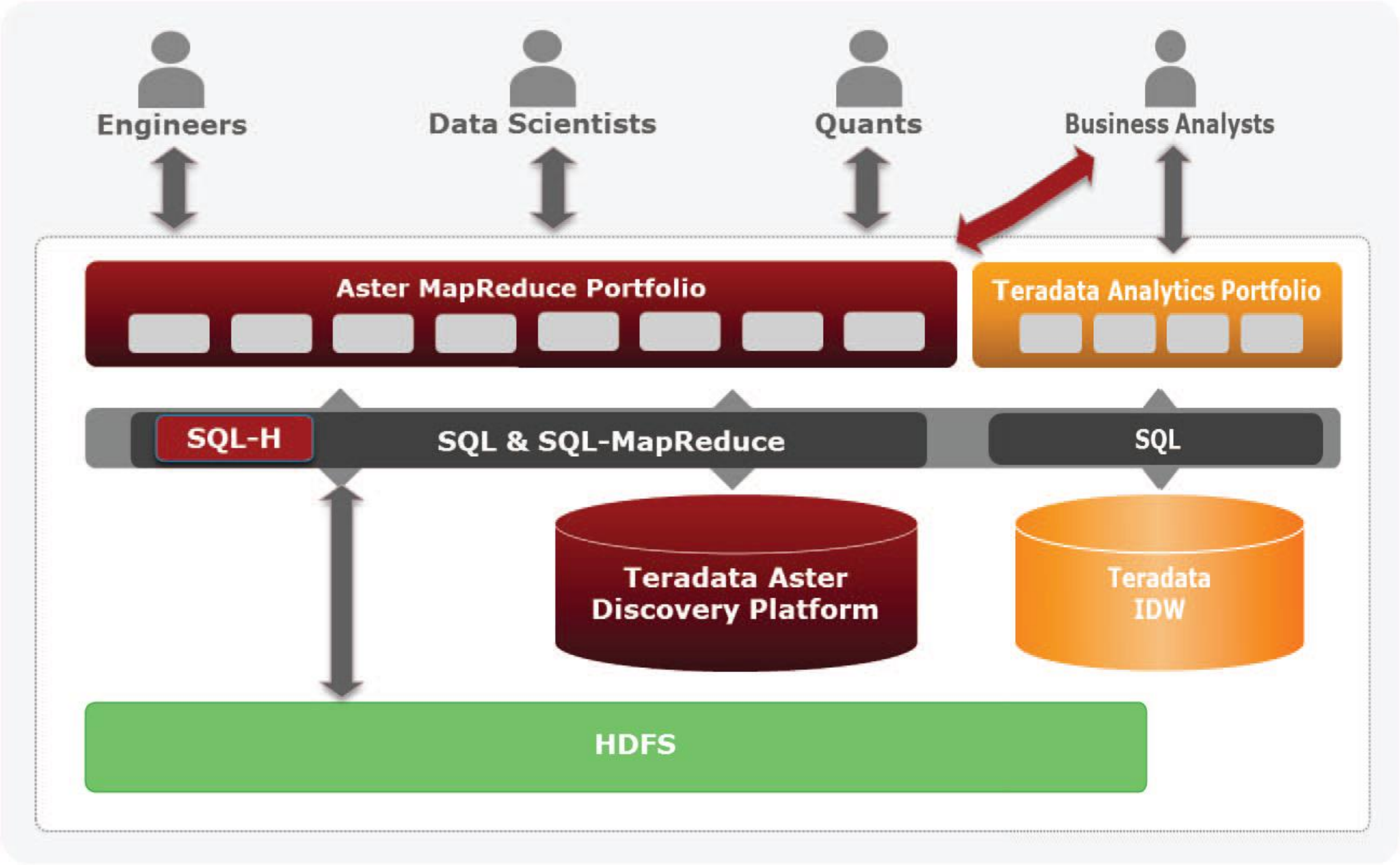
Aster Database Architecture

Aster Database not only stores large volumes of data but also processes data and analytic applications in-database to deliver faster, deeper insights


Aster Database
Time Series
Graph
Statistical
Other Aster-built functions
Partner-built functions
6
Pre-packaged advanced analytic frameworks
Unified Interface
SQL
SQL-MapReduce
5
Aster Data's SQL-MapReduce or SQL interface
High Volume, Fast Querying
Dynamic Workload Manager (WLM)
4
Industry-leading dynamic workload management
Massively-Parallel Data Stores
App
Data
Row Store
Column Store
3
Parallelized apps execute within the DBMS engine using MapReduce
2
Hybrid row/column MPP database for ultra-high performance
1
Commodity Hardware

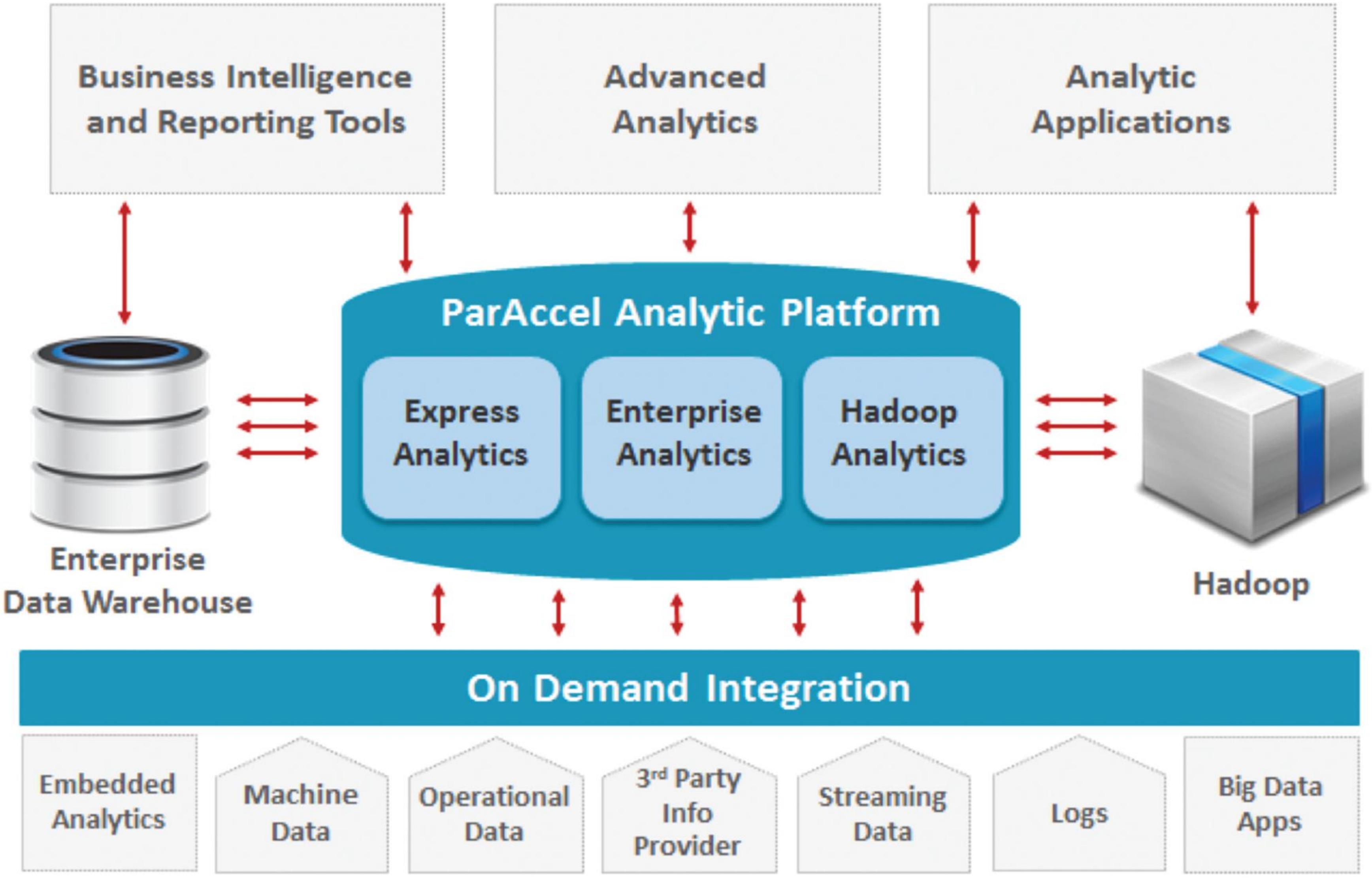
Advantage Teradata (+Aster database)

A combined view of the Teradata and Aster database suite



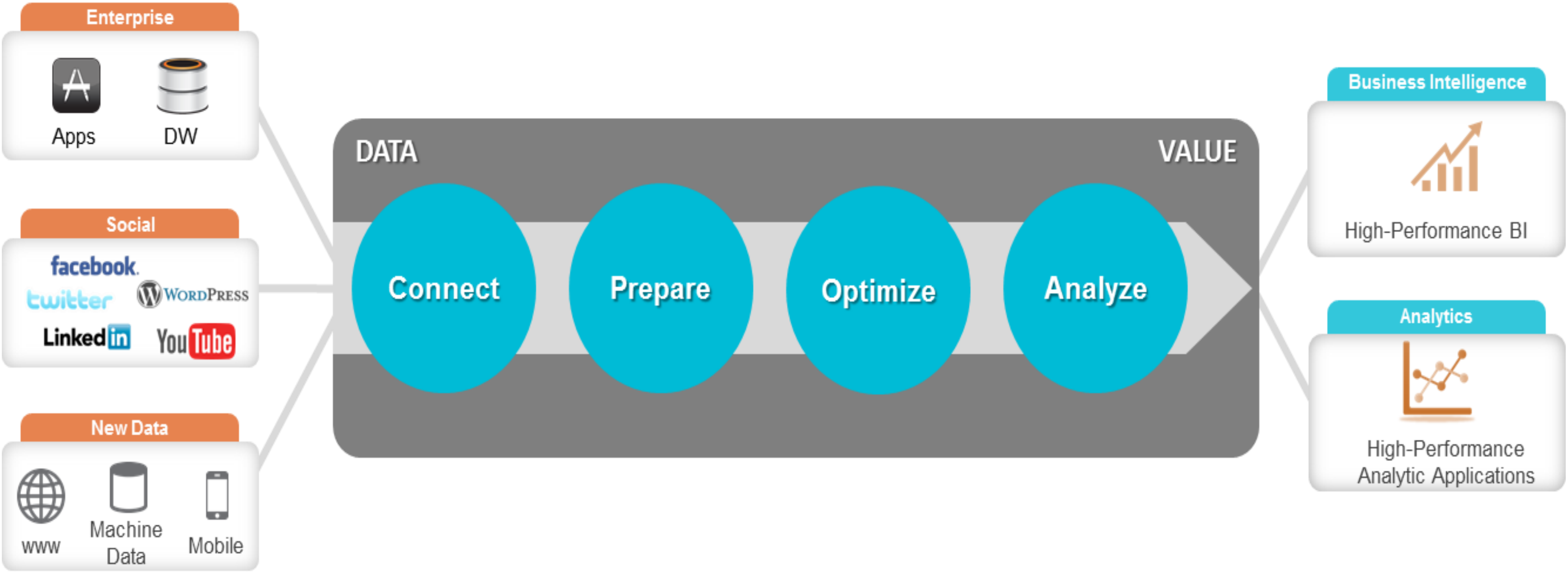
Paraccel

Paraccel Analytic Platform



Paraccel Hadoop Analytics

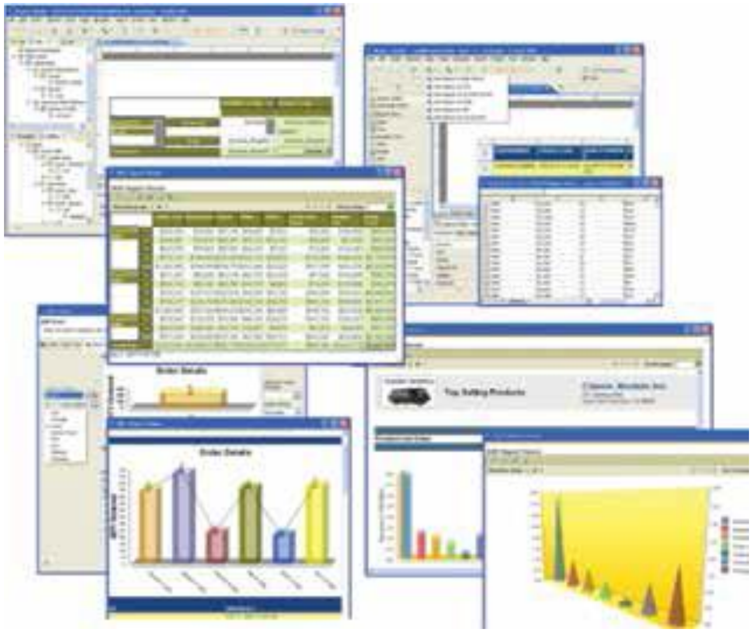
ParAccel Hadoop Analytics is a visual end-to-end data preparation and analytics model development and execution environment

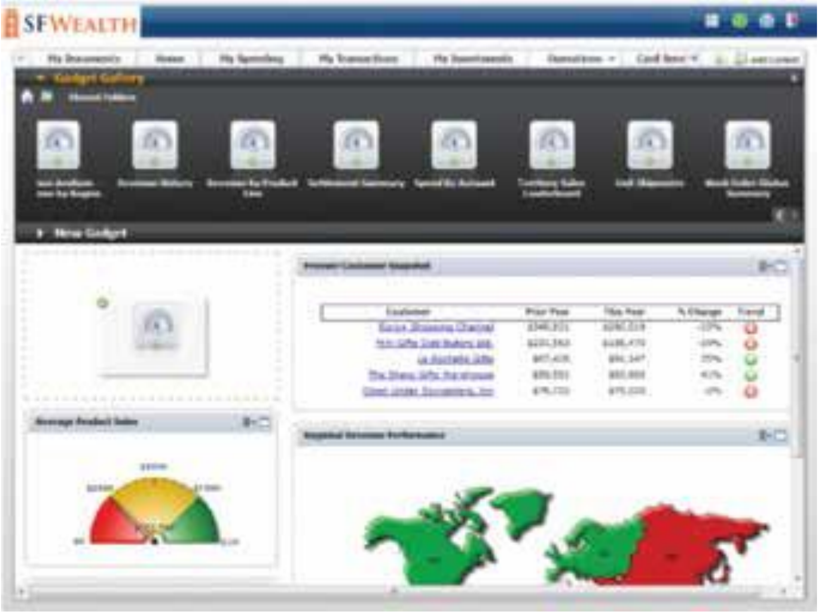


- Paraccel Hadoop Analytics automatically detects & utilizes all cores and nodes available at runtime up to a settable limit. It runs on any hardware, laptop, desktop, super-server or cluster

Paraccel Hadoop Analytics

- Display data graphically in interactive charts that can be modified on-the-fly. Scatter plots, line graphs, bar graphs, dashboards & more allow visualizing data to instantly spot trends. R provides a wide variety of open source options.





- Output through JDBC to visualization tools like Tableau, Actuate, YellowFin, etc. to expand visualization options without limits.

The Paraccel Edge

- Accelerated Data Preparation - Processing 250 million health insurance claims to find fraud and claims mismanagement took 48 days. With ParAccel scaling across all cores on all nodes, it took less than a day.

- Accelerated Analytics Design - Six MapReduce coders took a week to build a machine learning predictive analytics workflow on Hadoop. One analyst with ParAccel Hadoop Analytics reproduced the workflow in ten minutes, with more accurate predictive results, and not one line of code.

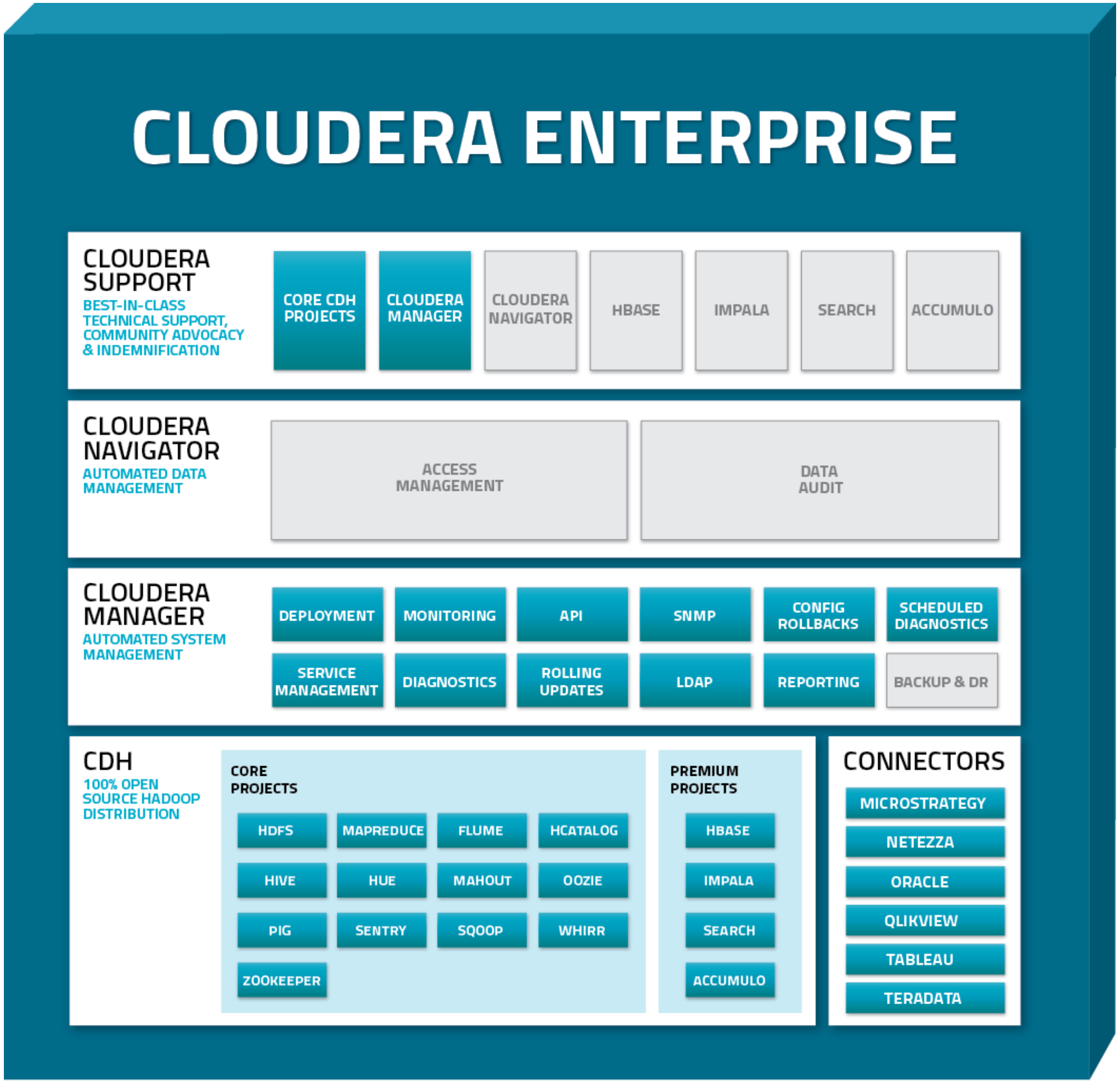
- Accelerated Execution Speed per Dollar - A risk management solution a global bank was using had a processing time of 15+ hours. ParAccel executed the same solution in20 minutes, on half the hardware.

- Accelerated Analytics Deployment - A global data science consultancy took a week or more for each customer when deploying custom risk management analytics applications. Using ParAccel, they regularly deploy those applications in a matter of hours.

Cloudera Enterprise

The leader in the Hadoop market with 100+ customers

- Adds Cloudera Manager – Console for administering & managing Hadoop deployments and Enterprise Support
- Cloudera Manager offers wizard-based installation and configuration menus for deploying Hadoop. Also offers tools to help system managers monitor the health of the platform, diagnose problems, optimize performance, and make required configuration and security changes



CLOUDERA ENTERPRISE

CLOUDERA SUPPORT
BEST-IN-CLASS TECHNICAL SUPPORT, COMMUNITY ADVOCACY & INDEMNIFICATION

CORE CDH PROJECTS | CLOUDERA MANAGER | CLOUDERA NAVIGATOR | HBASE | IMPALA | SEARCH | ACCUMULO

CLOUDERA NAVIGATOR
AUTOMATED DATA MANAGEMENT

ACCESS MANAGEMENT | DATA AUDIT

CLOUDERA MANAGER
AUTOMATED SYSTEM MANAGEMENT

DEPLOYMENT | MONITORING | API | SNMP | CONFIG ROLLBACKS | SCHEDULED DIAGNOSTICS
SERVICE MANAGEMENT | DIAGNOSTICS | ROLLING UPDATES | LDAP | REPORTING | BACKUP & DR

CDH
100% OPEN SOURCE HADOOP DISTRIBUTION

CORE PROJECTS
HDFS | MAPREDUCE | FLUME | HCATALOG
HIVE | HUE | MAHOUT | OOZIE
PIG | SENTRY | SQOOP | WHIRR
ZOOKEEPER

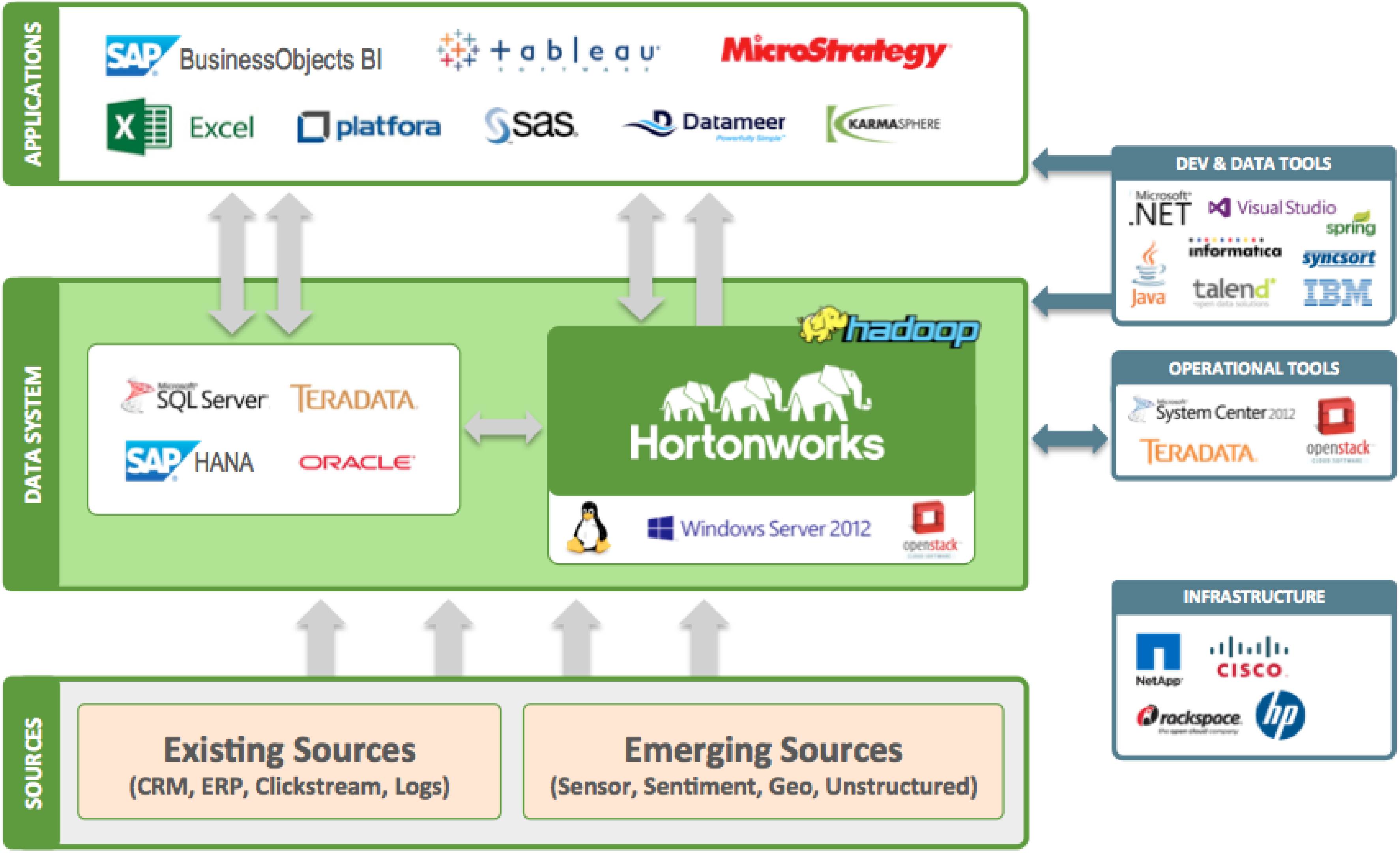
PREMIUM PROJECTS
HBASE
IMPALA
SEARCH
ACCUMULO

CONNECTORS
MICROSTRATEGY
NETEZZA
ORACLE
QLIKVIEW
TABLEAU
TERADATA

🟦 CLOUDERA ENTERPRISE

TLS

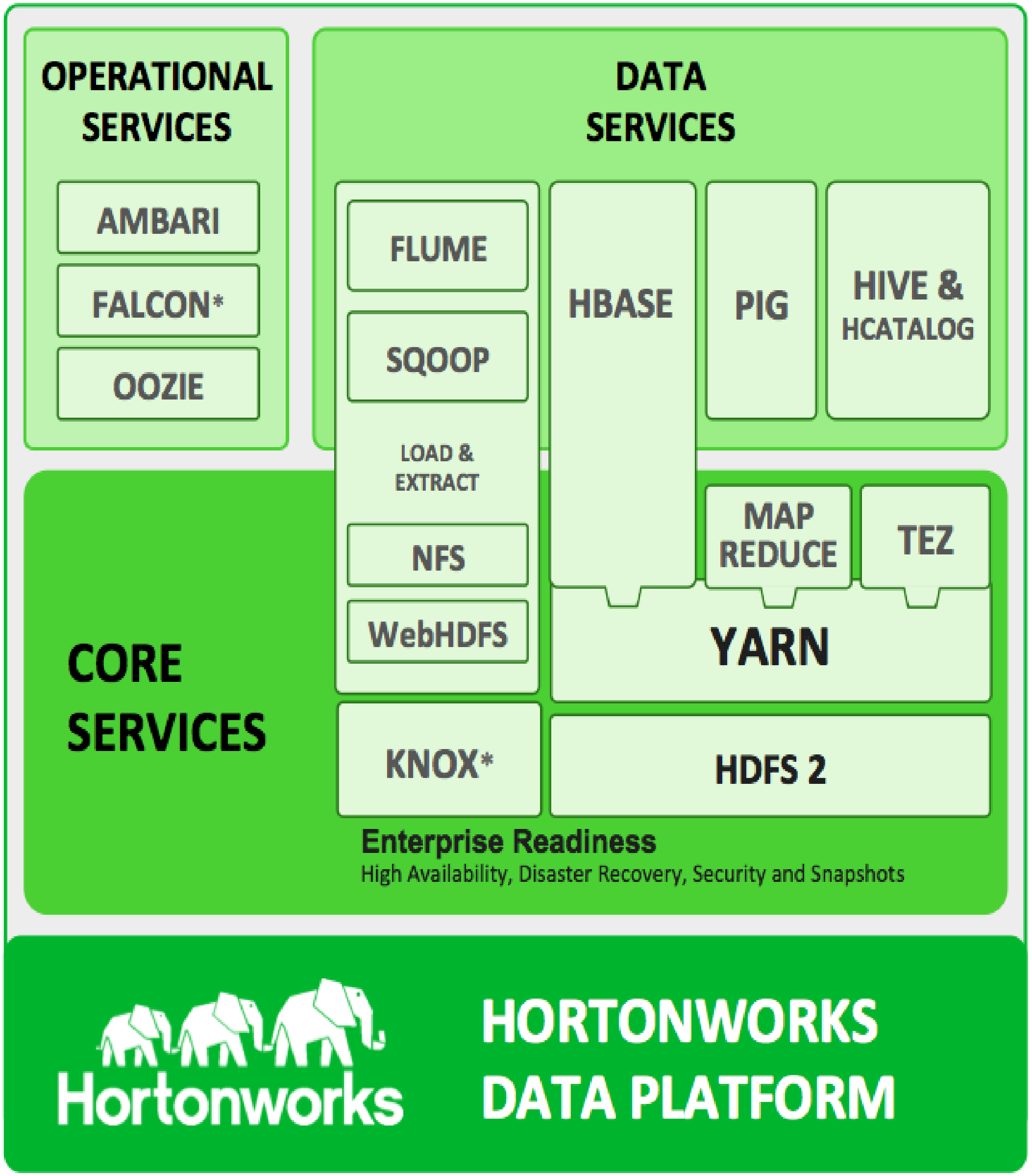
HortonWorks



TLS

HortonWorks

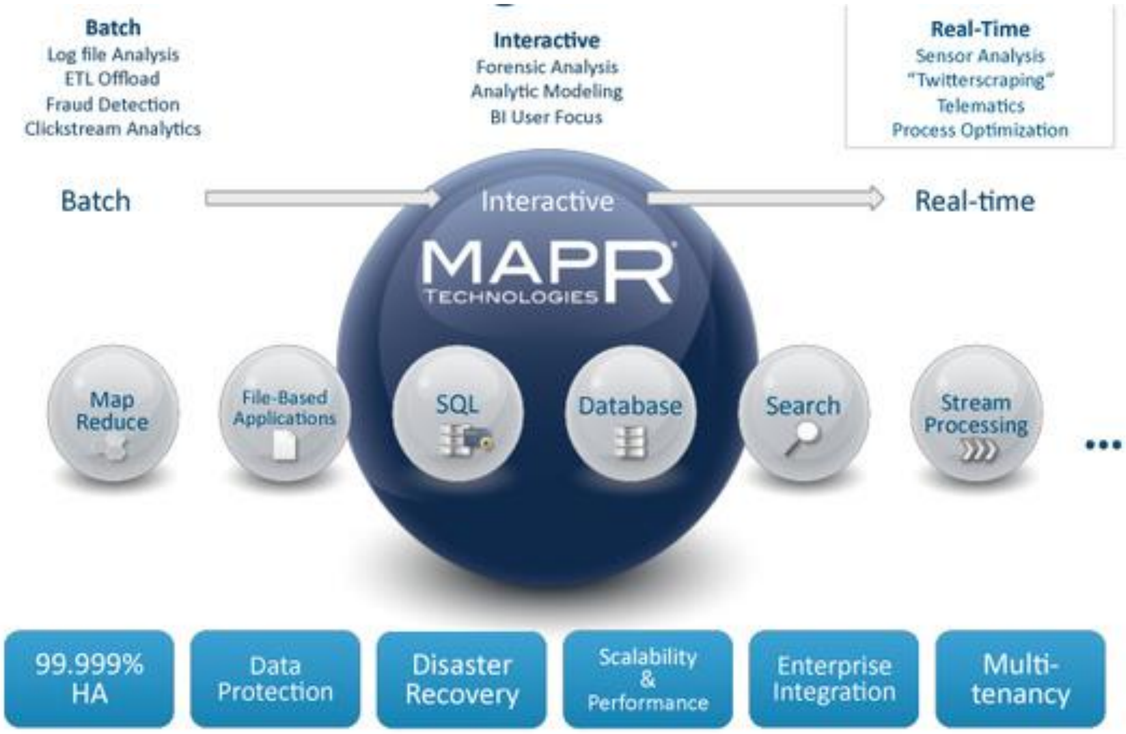
- Comes in 3 editions –
  - HDP (2 or 1.3)
  - HDP for Windows
  - HDP Sandbox
- Include Apache Ambari (similar to Cloudera Manager)
- BCP improvements via snapshots for recovery
- Better connectors to Oracle, Netezza
- Improved functionality and performance in Sqoop
- Improve SQL support using Hive



TLS

MapR

Arguably the strongest challenger to Cloudera

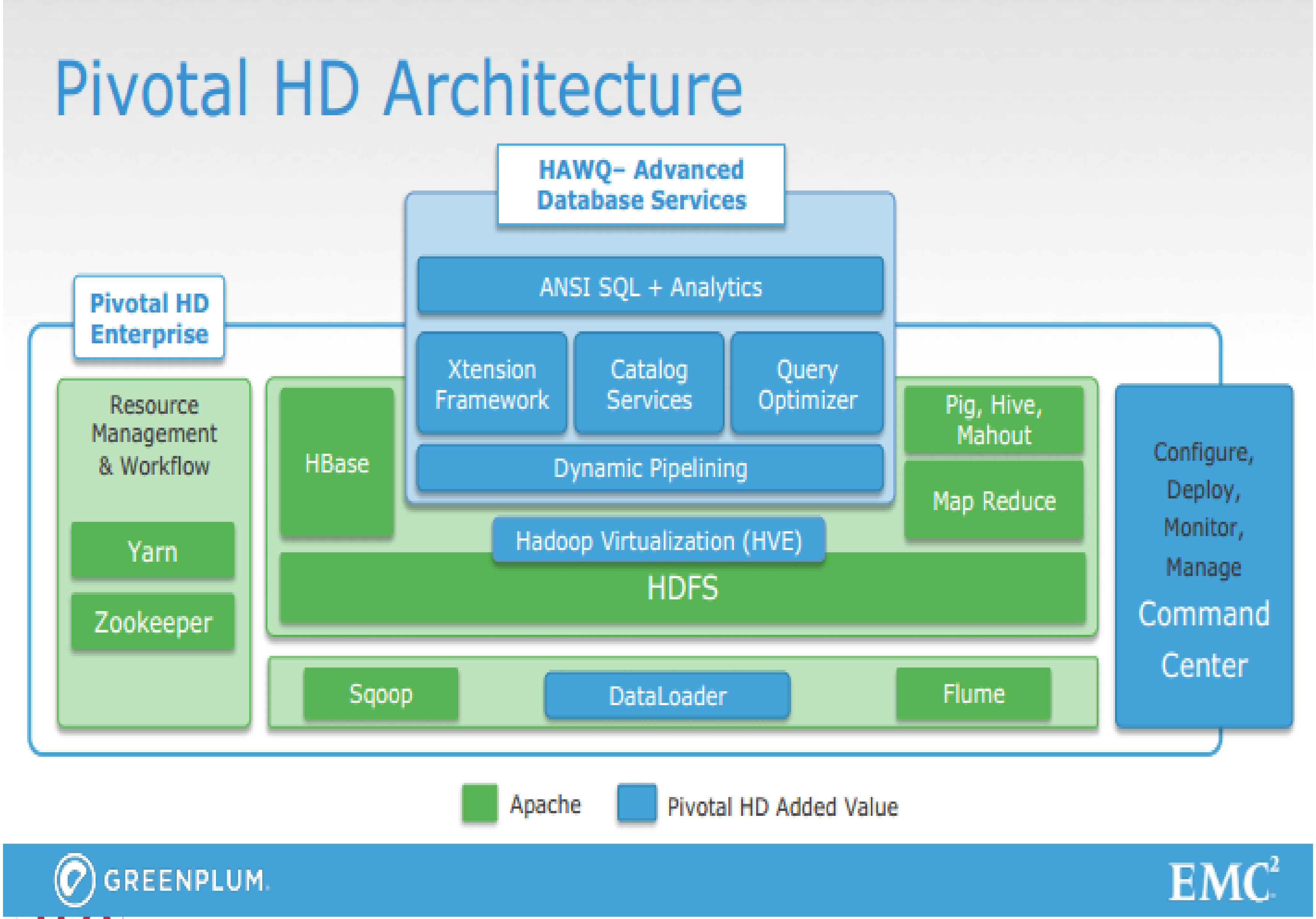


- Comes in 3 editions – M3, M5 and M7
- M3 is Free Hadoop, M5 is Enterprise Hadoop and M7 Enterprise NoSQL (Hbase) and Hadoop
- Proprietary Mountable file system
- High availability name nodes & job trackers

- Amazon’s Elastic MapReduce now includes MapR M7
- Was the underlying Hadoop distro used in EMC Greenplum (called Greenplum HD)

TLS

EMC Greenplum’s Pivotal



Pivotal HD Architecture

HAWQ- Advanced Database Services

ANSI SQL + Analytics

Xtension Framework

Catalog Services

Query Optimizer

Dynamic Pipelining

Pivotal HD Enterprise

Resource Management & Workflow

Yarn

Zookeeper

HBase

Pig, Hive, Mahout

Map Reduce

Hadoop Virtualization (HVE)

HDFS

Sqoop

DataLoader

Flume

Configure, Deploy, Monitor, Manage

Command Center

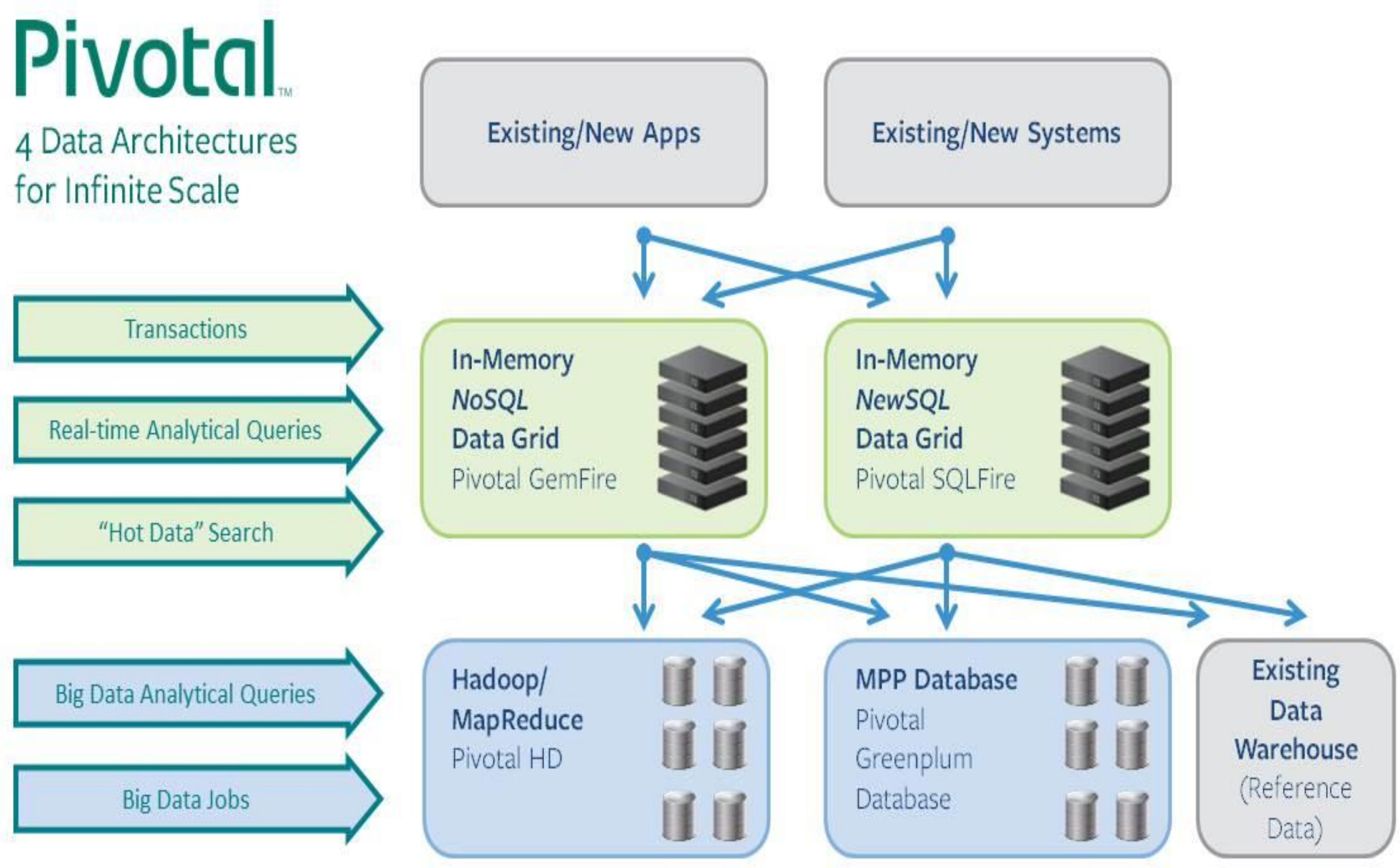
🟩 Apache 🟦 Pivotal HD Added Value

GREENPLUM

EMC²

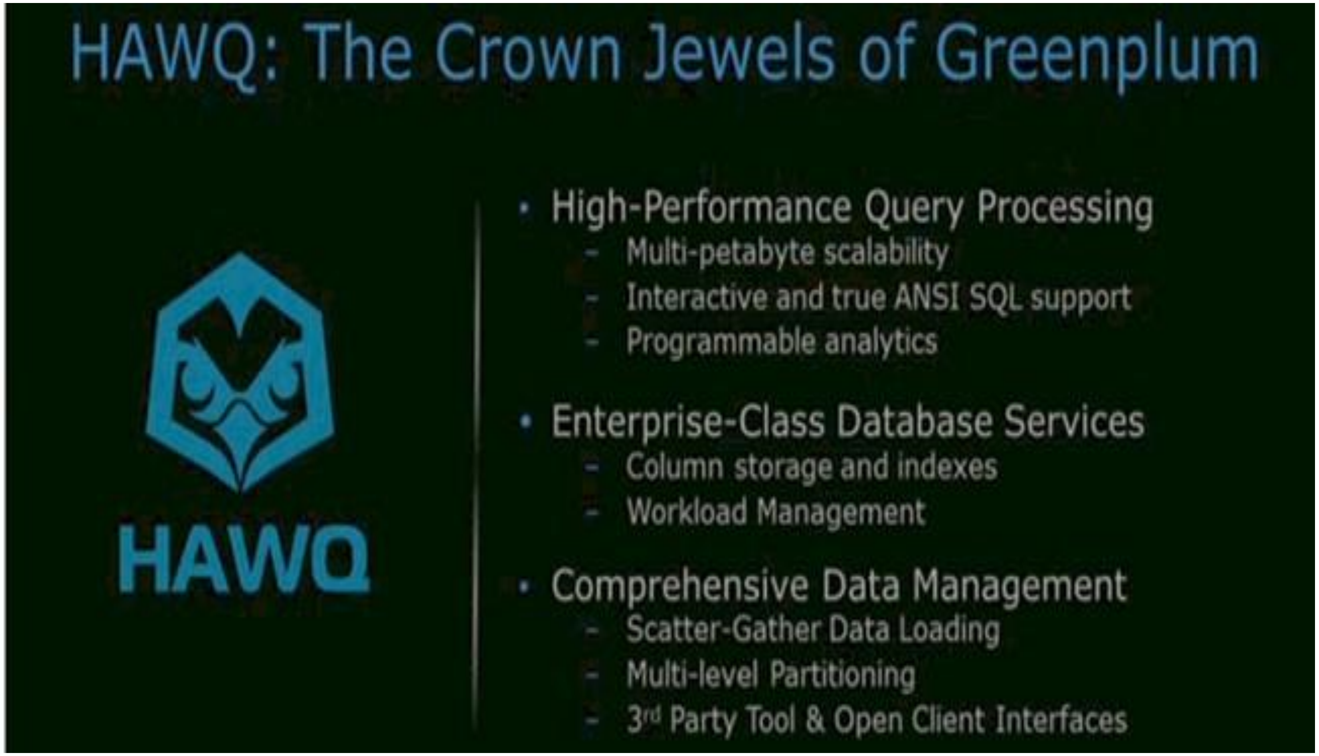
Pivotal Analytics

Pivotal analytics is a unique end-to-end analytics platform that fills the need for an elastically scalable, real-time analytics solution that delivers big answers.



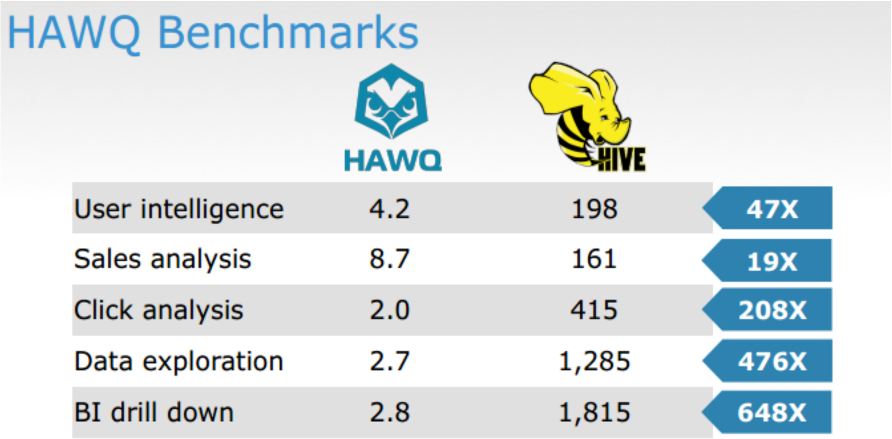
Pivotal™
4 Data Architectures for Infinite Scale
Existing/New Apps
Existing/New Systems
Transactions
Real-time Analytical Queries
"Hot Data" Search
In-Memory NoSQL Data Grid Pivotal GemFire
In-Memory NewSQL Data Grid Pivotal SQLFire
Big Data Analytical Queries
Big Data Jobs
Hadoop/ MapReduce Pivotal HD
MPP Database Pivotal Greenplum Database
Existing Data Warehouse (Reference Data)

Greenplum HD Pivotal – HAWQ



HAWQ: The Crown Jewels of Greenplum

- High-Performance Query Processing
  - Multi-petabyte scalability
  - Interactive and true ANSI SQL support
  - Programmable analytics
- Enterprise-Class Database Services
  - Column storage and indexes
  - Workload Management
- Comprehensive Data Management
  - Scatter-Gather Data Loading
  - Multi-level Partitioning
  - 3rd Party Tool & Open Client Interfaces



HAWQ Benchmarks

	HAWQ	HIVE	
User intelligence	4.2	198	47X
Sales analysis	8.7	161	19X
Click analysis	2.0	415	208X
Data exploration	2.7	1,285	476X
BI drill down	2.8	1,815	648X

TLS

Bigdata Use Cases

Bigdata Use Cases – HealthCare

HealthCare

- Clinical Decision Support System
- Remote Patient Monitoring
- Patient Profiling
- Comparative Effective Research

TLS

Bigdata Use Cases – Retail

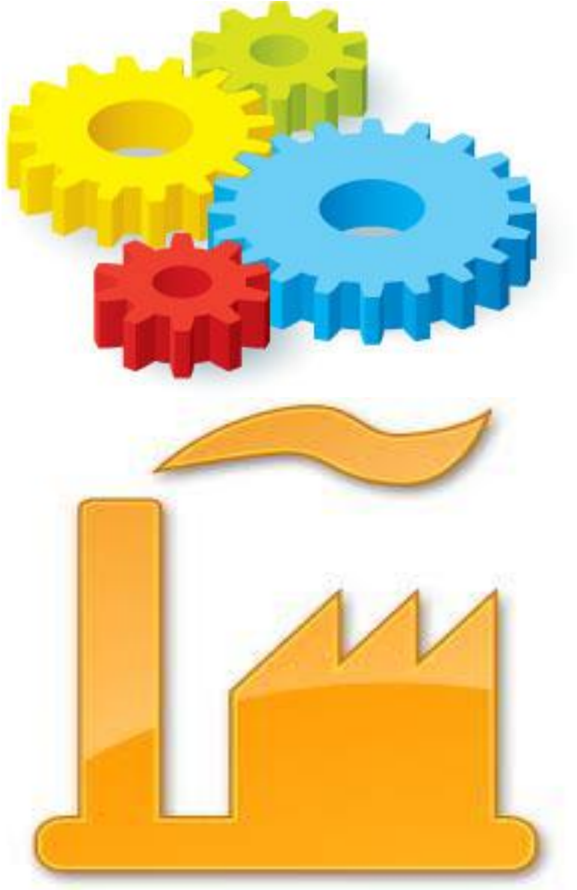
Marketing

- Cross-Selling
- Location-based Marketing
- Analyzing in – store behavior
- Sentiment Analysis

Bigdata Use Cases – Manufacturing

Manufacturing

- Sensors Driven Operations
- Supply Chain & Inventory Management
- Shorten Design to Value Cycle



TLS

Bigdata Use Cases – Public Services

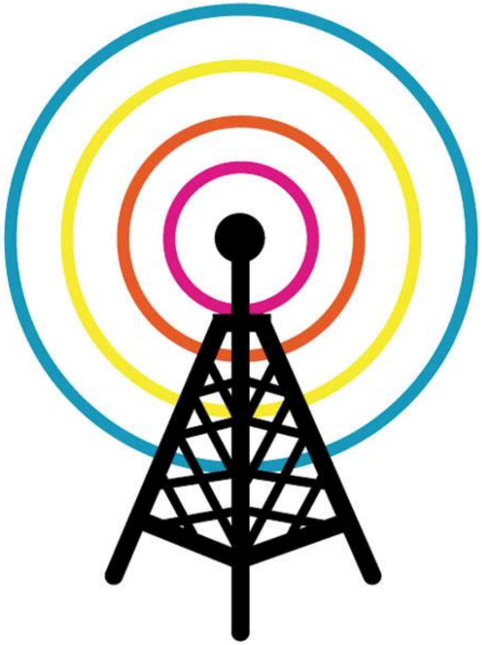
Public Sector Administration

- Traffic Decongestion
- Civic Compliance – Noise, Air & Water Pollution
- Locality and Area Development

Bigdata Use Cases – Telecom

Telecom

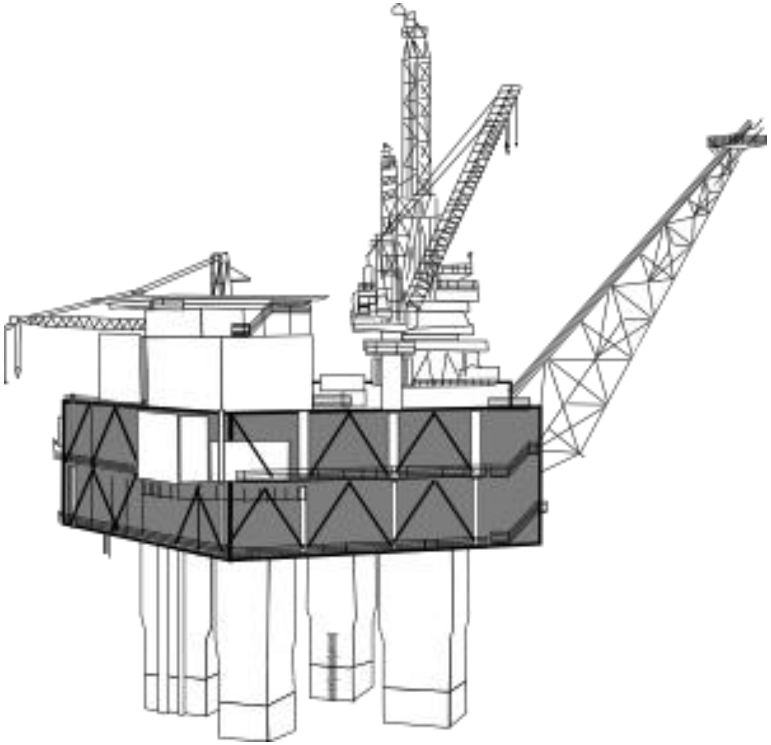
- Geo – Targeted Advertising
- Electronic Toll Collection
- Emergency Response
- Remote Personal Car safety & monitoring
- Urban Planning



Bigdata Use Cases – Oil & Gas

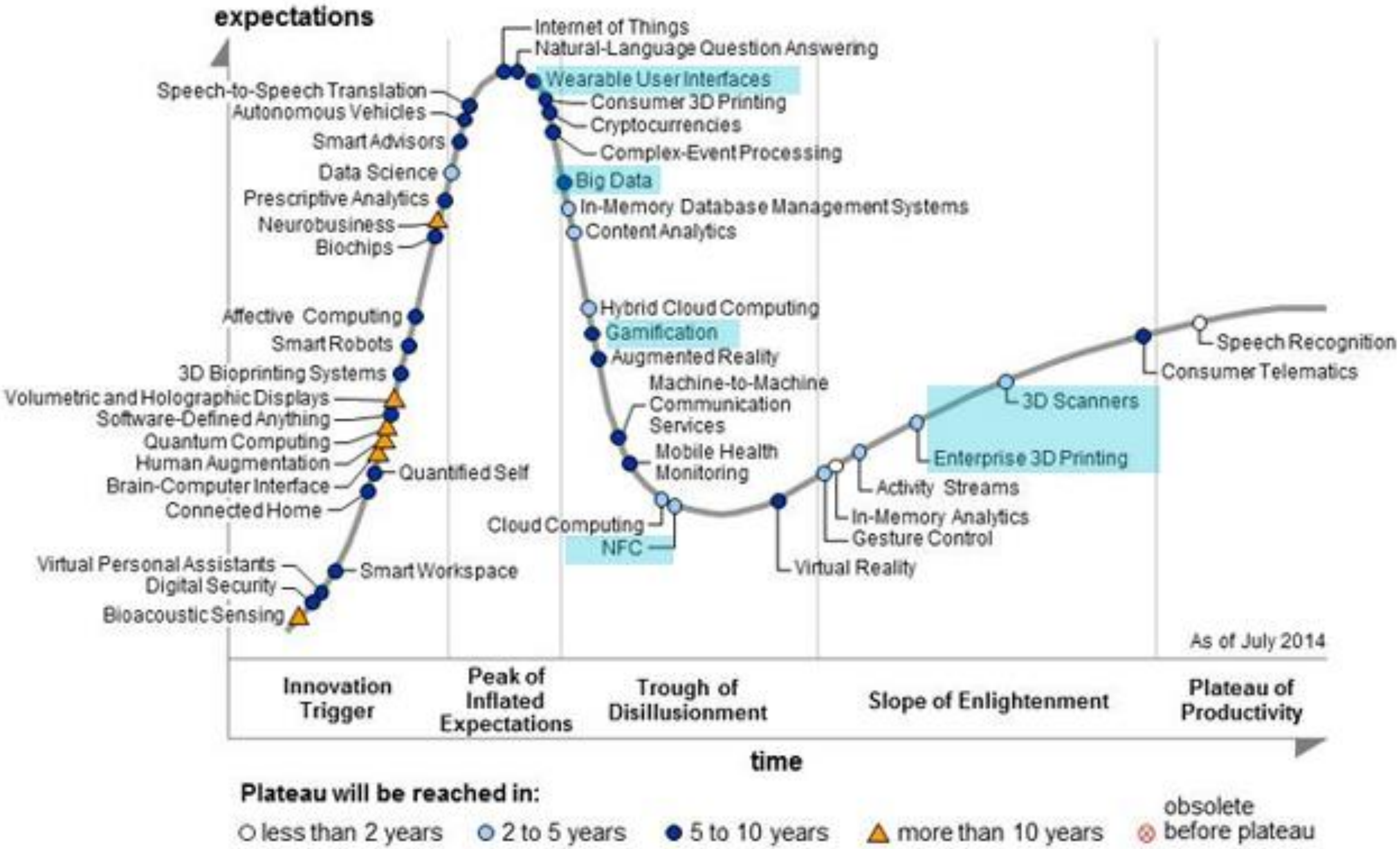
Oil & Gas Sector

- Digital Oil Field
- Remote Drilling Operations



The Bigdata Market

Gartner Hype Cycle on Bigdata



Some Unanswered Questions

- Too much emphasis on the 3Vs & 1C, what about Value?
- Conventional BI has mature and elaborate processes and definitions for data validation, data integrity, data governance, data extraction, metadata management and ETL, what about these disciplines in Bigdata?
- Convergence of Structured and Unstructured data (single query) how will it be handled?
- Challenges with data update and data erasure
- Writing MapReduce in Java is cumbersome and inefficient, need quick evolution of IDEs and GUIs
- Convergence of BI capability & Java programming – Cross skilling, multi-skilling challenges

Thank You

Disclaimer

Tech Mahindra Limited, herein referred to as TechM provide a wide array of presentations and reports, with the contributions of various professionals. These presentations and reports are for informational purposes and private circulation only and do not constitute an offer to buy or sell any securities mentioned therein. They do not purport to be a complete description of the markets conditions or developments referred to in the material. While utmost care has been taken in preparing the above, we claim no responsibility for their accuracy. We shall not be liable for any direct or indirect losses arising from the use thereof and the viewers are requested to use the information contained herein at their own risk. These presentations and reports should not be reproduced, re-circulated, published in any media, website or otherwise, in any form or manner, in part or as a whole, without the express consent in writing of TechM or its subsidiaries. Any unauthorized use, disclosure or public dissemination of information contained herein is prohibited. Unless specifically noted, TechM is not responsible for the content of these presentations and/or the opinions of the presenters. Individual situations and local practices and standards may vary, so viewers and others utilizing information contained within a presentation are free to adopt differing standards and approaches as they see fit. You may not repackage or sell the presentation. Products and names mentioned in materials or presentations are the property of their respective owners and the mention of them does not constitute an endorsement by TechM. Information contained in a presentation hosted or promoted by TechM is provided “as is” without warranty of any kind, either expressed or implied, including any warranty of merchantability or fitness for a particular purpose. TechM assumes no liability or responsibility for the contents of a presentation or the opinions expressed by the presenters. All expressions of opinion are subject to change without notice.