Big Data & Its Bigger Possibilities In The Cloud

Chhavi Gupta
Software Engineer, EMC Corporation

Sai Pattem
Professional MBA Candidate 2013

August 15th , 9:30 – 10:30 AM
Session – 13860
Room 200 (Hynes Convention Center)

EMC2

Agenda

- Big Data
  - Definition
  - Relativity
  - Challenges
- Cloud Computing
  - Definition
  - Private, Public, and Hybrid Cloud
  - SaaS
  - PaaS
  - IaaS
- Technology
- Tools
  - Handling Big Data in the Cloud
- Conclusion

Big Data

- Definition of Big Data consists of 3Vs+C
  - High Volume (Facebook,youTube)
  - High Velocity (Facebook,Twitter)
  - High Variety (text files, multimedia, pdfs)
  - Complexity (Amazon)

Mobile Sensors

Social Media

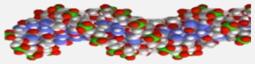
Video Surveillance

Video Rendering

Smart Grids

Geophysical Exploration

Medical Imaging



Gene Sequencing

Big Data

Relativity

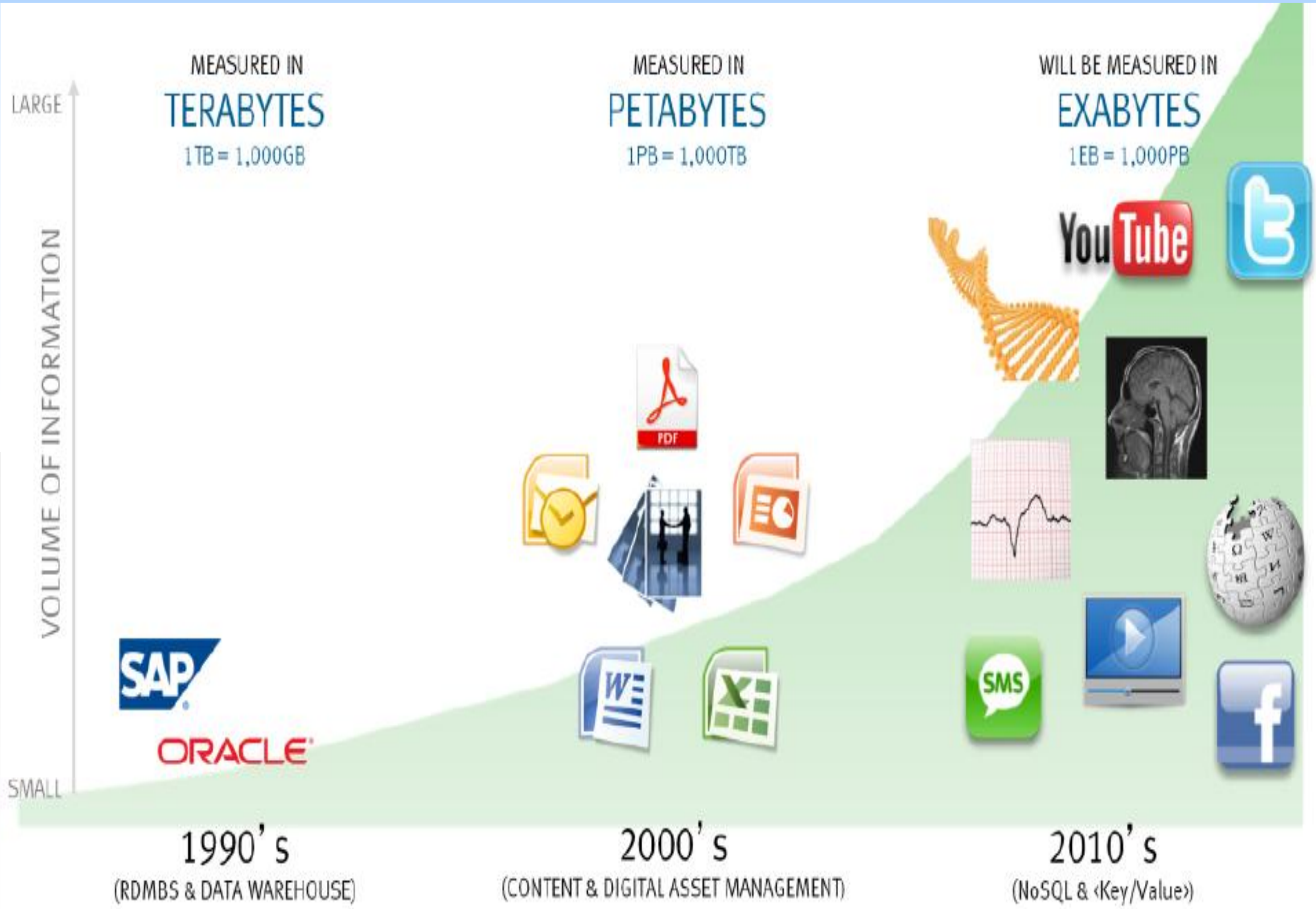
Big Data is a relative concept

What is BIG today…

May not be so big tomorrow….

Big Data

Relativity



Big Data

Challenges

- ➢ Challenges related to big data
  - – Organization needs to grow but can’t spend much to buy new servers, storage
  - – Reliable backup and need to access anywhere/anytime
  - – Want to test a software before investment in it
  - – May need an application for only a brief period of time
  - – Critical customer data, but lacks secured storage infrastructure

Cloud

Introduction

- What is Cloud Computing?
  - ✓ Massively scalable
  - ✓ Convenient on-demand network access
  - ✓ Enables an organization to extend virtualization beyond enterprise data center
  - ✓ Aggregates resources scattered across the globe
  - ✓ Location independent virtual image of aggregated resources
  - ✓ Fully- automated request fulfillment process in the background

Cloud

Pros & Cons

Private
Public
Hybrid

Private Cloud

1) Higher Security
2) Higher Control
3) Better Service Quality
4) Higher Availability

1) More Maintenance
2) Big or Mid-size Companies

Public Cloud

1) Cost Efficient
2) Competitive Advantage
3) Readily Available

1) Less Secured & Compliant
2) Higher Data Vulnerabilities

Hybrid Cloud

1) Combine Multiple Services to Increase Overall Capability or Capacity
2) Improved Resiliency and Disaster Recovery
3) Better Service Quality
4) Complex Architectural and Design Needs

EMC²

Cloud

SaaS

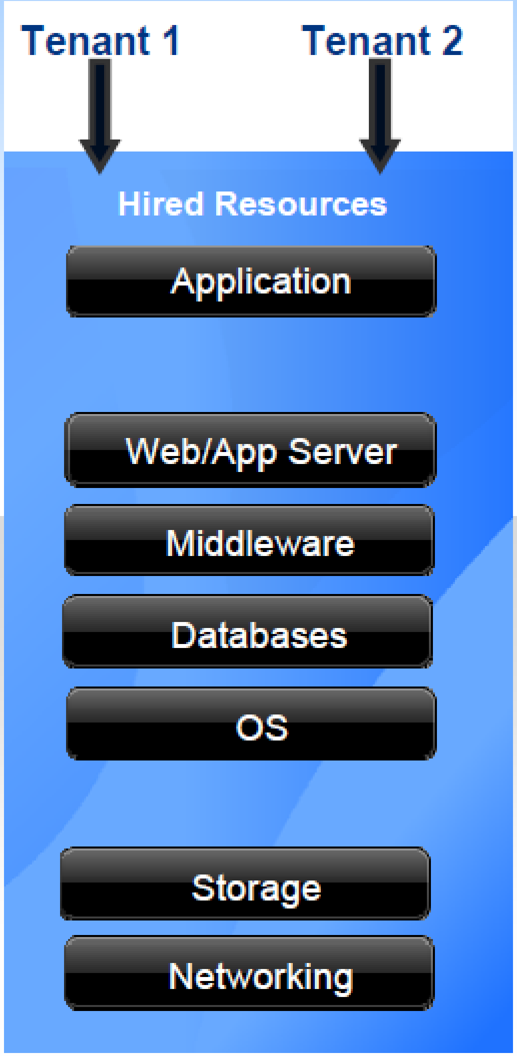
SaaS
Software-as-a-Service

➢ **You pay for the application**

- Apps accessible from various client devices
- Through a web browser

For example:

- Salesforce.com
- EMC Mozy (Backup as the service)
- Google Apps



[Source: NIST]

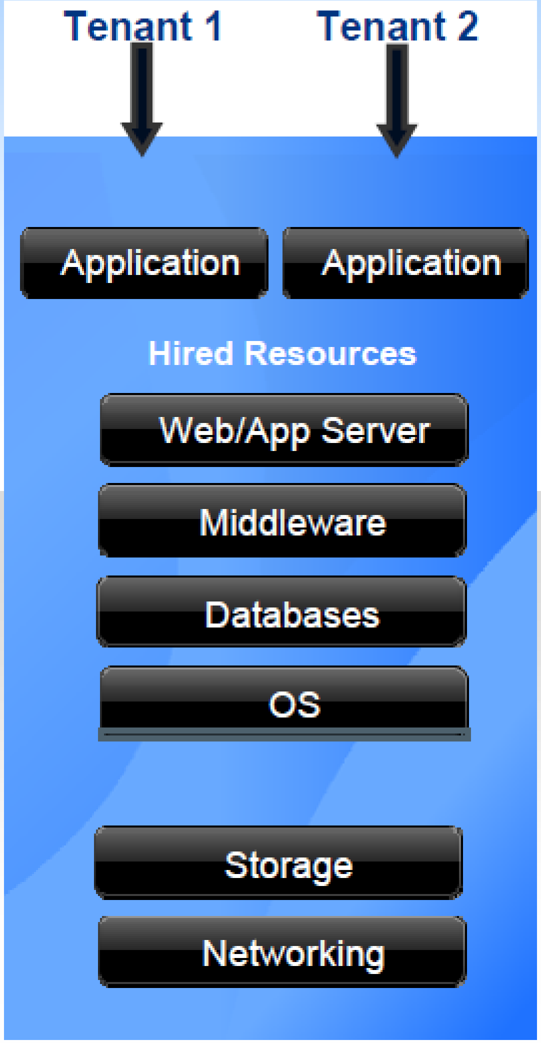
PaaS
Platform-as-a-Service

- ➢ You pay for the platform software components
- ➢ Your applications are built on top

For example:

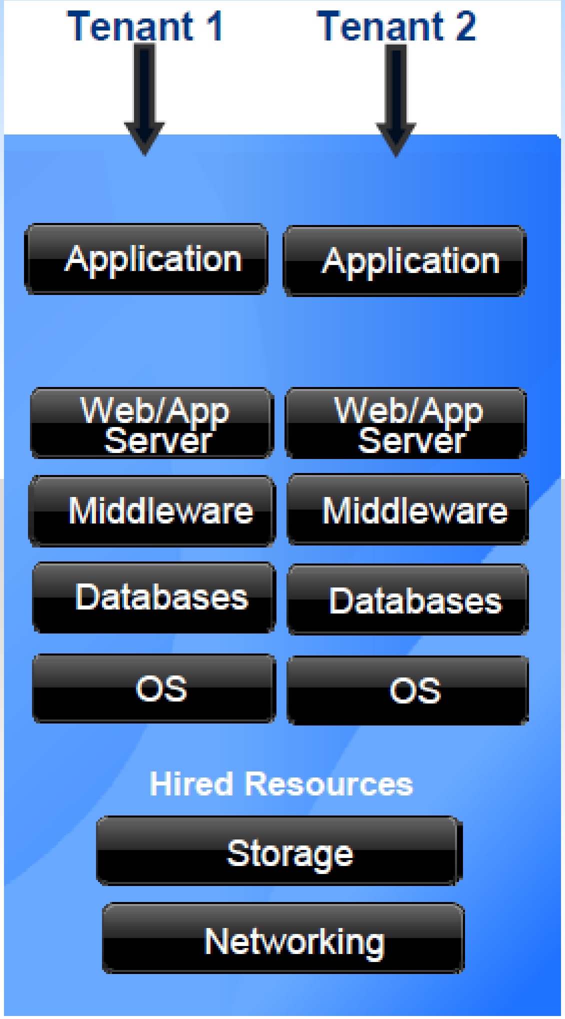
- – Google App Engine
- – Microsoft Azure
- – Force.com Platform

[Source: NIST]



IaaS
Infrastructure-as-a-Service

- ➢ You pay for the infrastructure components
- ➢ Your OS image and applications on top



For example:

- – Amazon EC2
- – EMC Atmos

[Source: NIST]

Technology

Big Data & Cloud

TECHNIQUES

+

TECHNOLOGIES

Handling Data

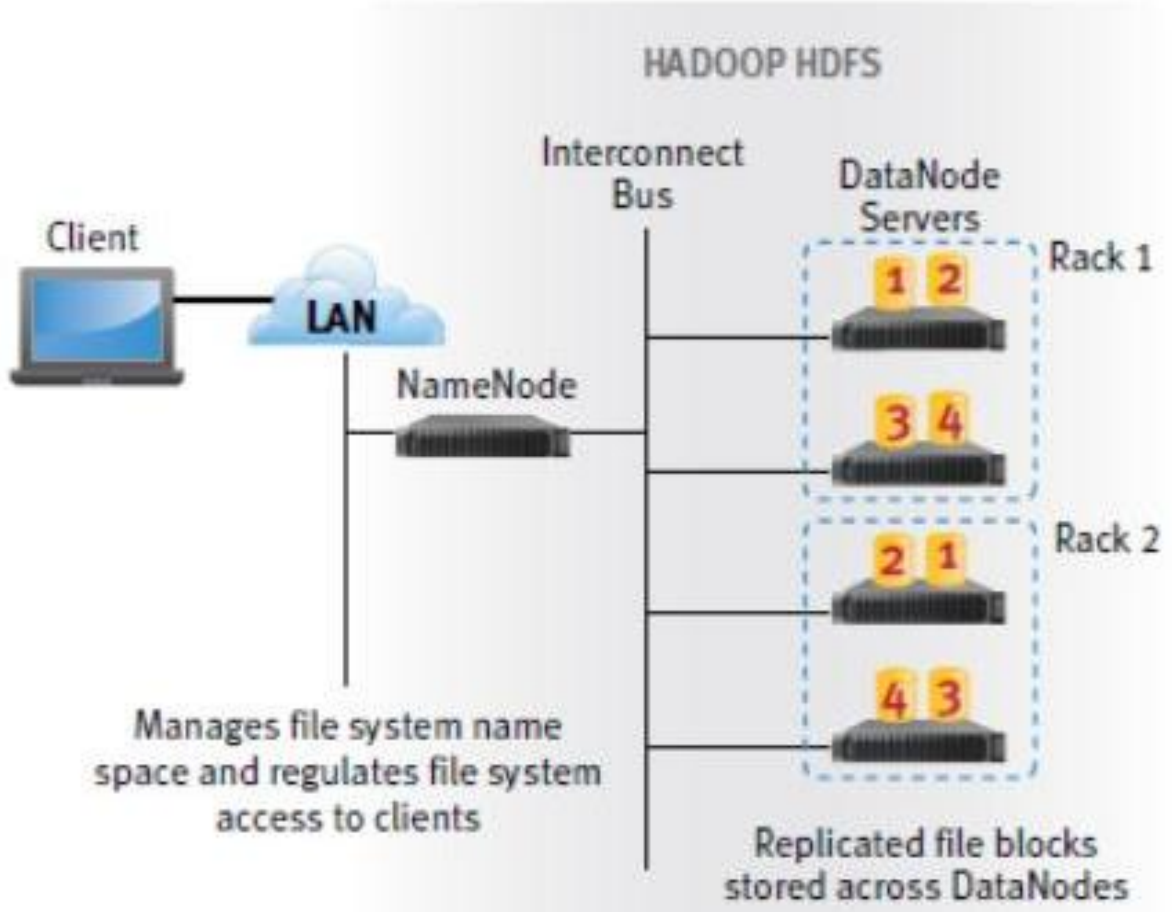
Extreme Scale

AFFORDABLE

Forrester Research

EMC²

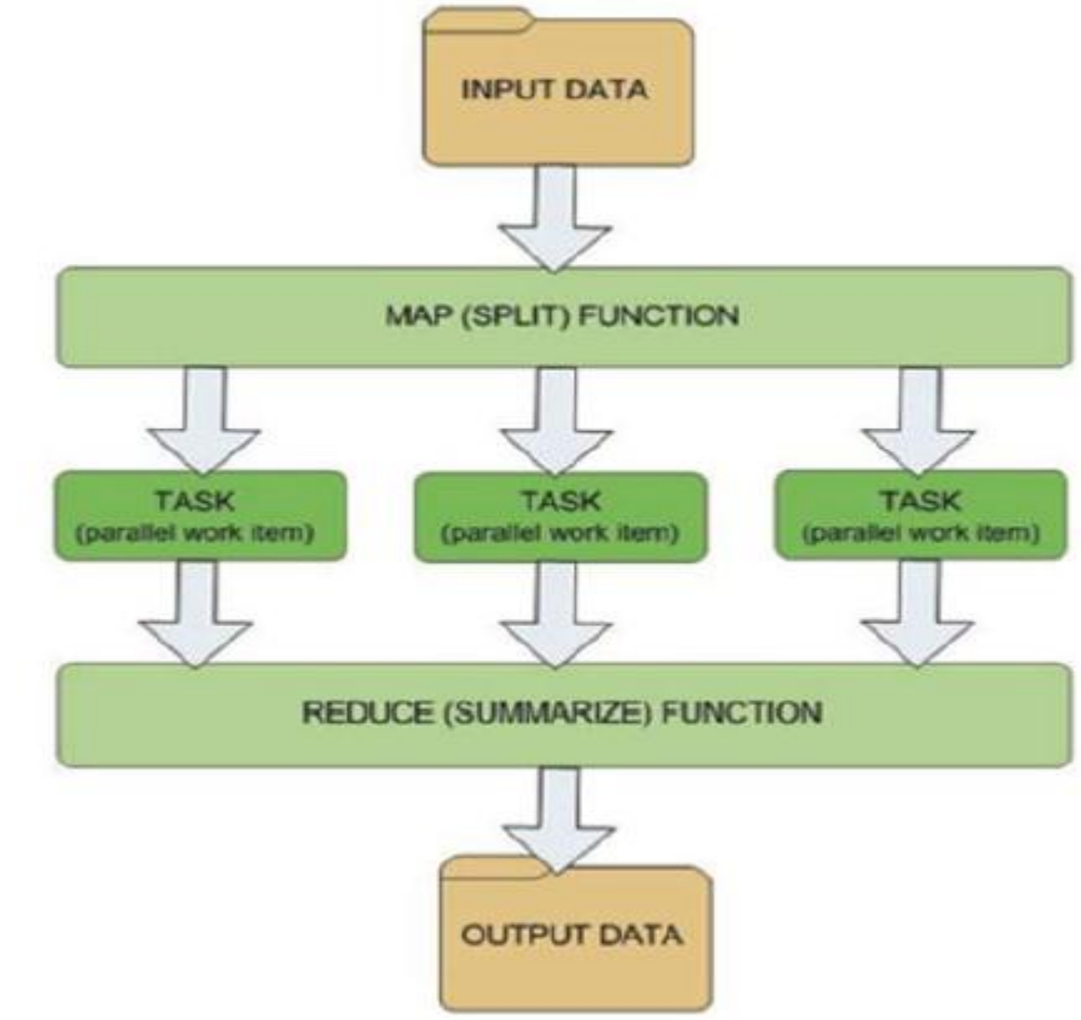
HDFS: Hadoop Distributed File System


HADOOP HDFS
Interconnect Bus
DataNode Servers
Client
LAN
Rack 1
1 2
3 4
NameNode
Rack 2
2 1
4 3
Manages file system name space and regulates file system access to clients
Replicated file blocks stored across DataNodes

Technology

Big Data & Cloud

Basic Map/Reduce Data Flow



INPUT DATA

MAP (SPLIT) FUNCTION

TASK (parallel work item)

TASK (parallel work item)

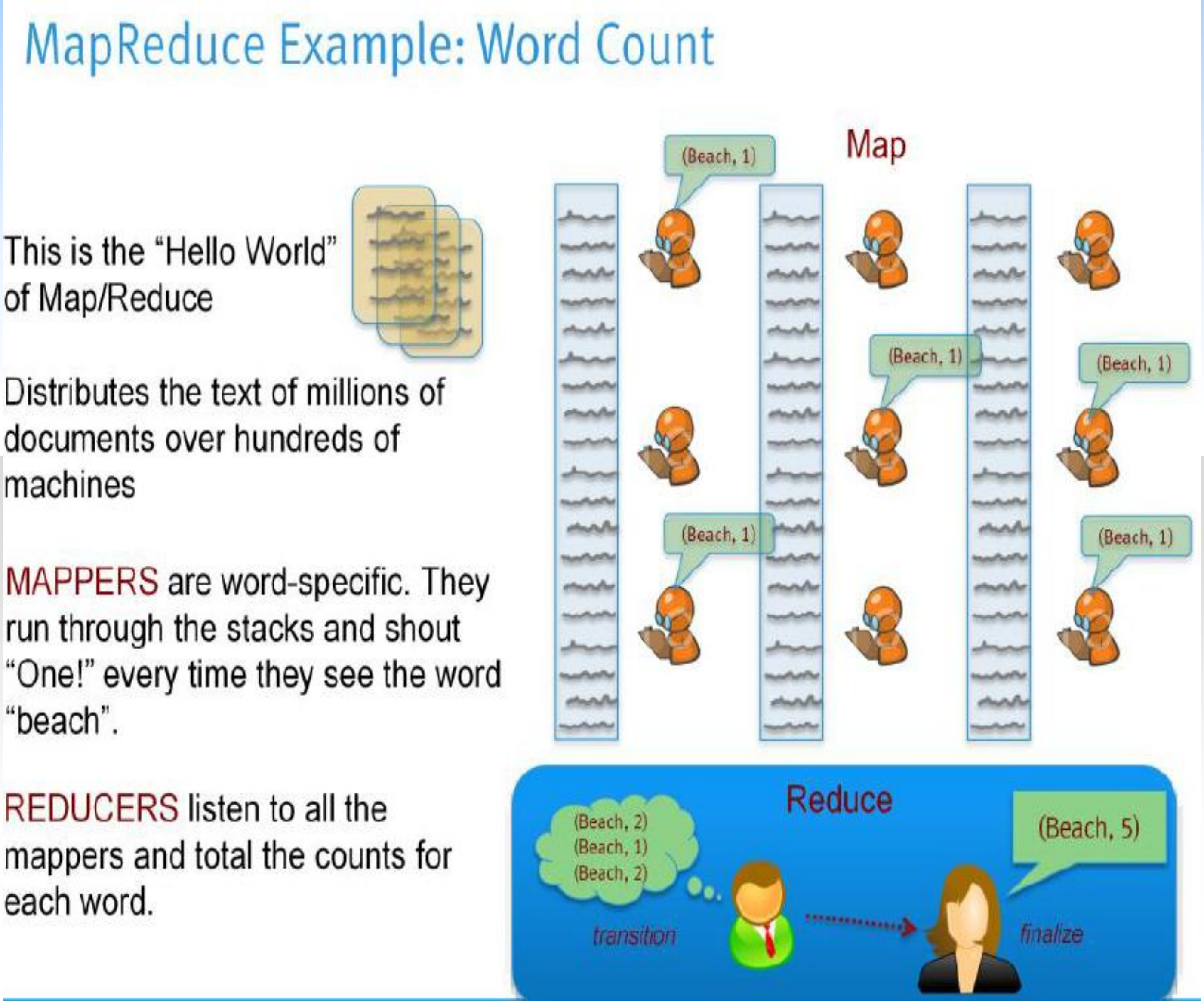
TASK (parallel work item)

REDUCE (SUMMARIZE) FUNCTION

OUTPUT DATA

EMC²

Big Data & Cloud



MapReduce Example: Word Count

This is the “Hello World” of Map/Reduce

Distributes the text of millions of documents over hundreds of machines

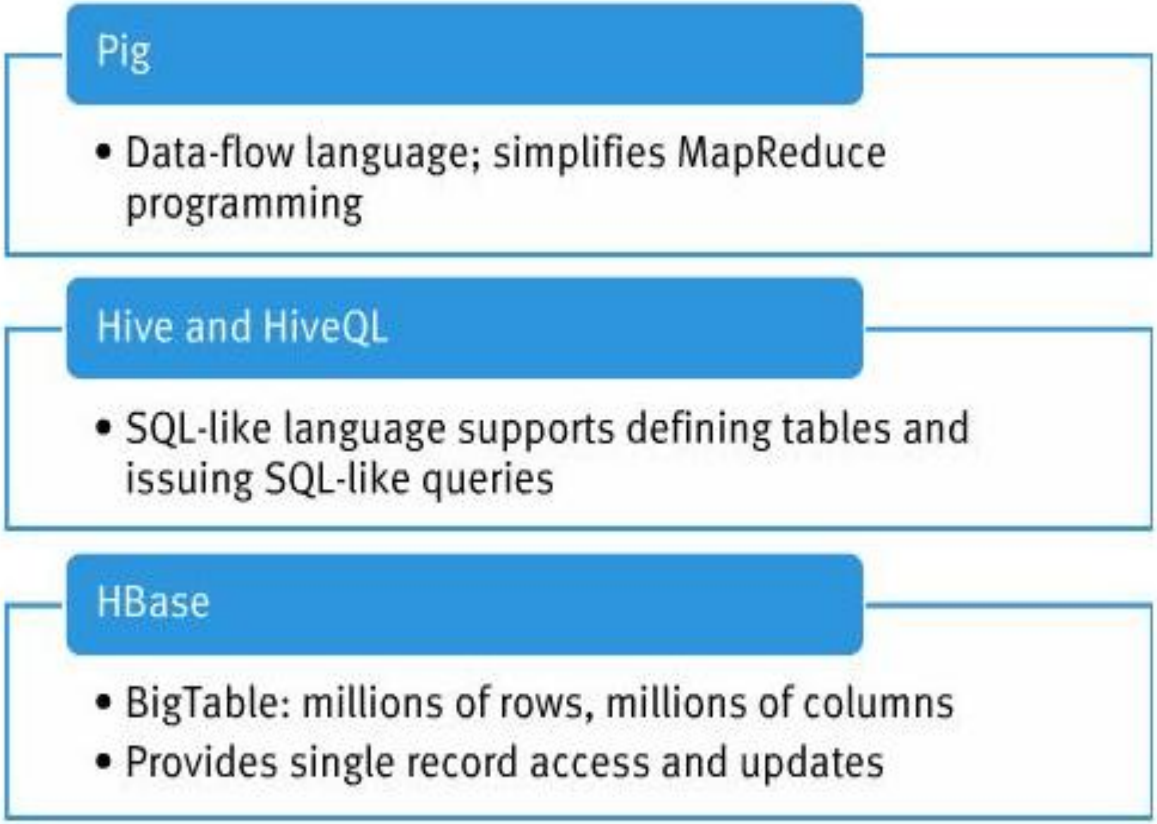
MAPPERS are word-specific. They run through the stacks and shout “One!” every time they see the word “beach”.

REDUCERS listen to all the mappers and total the counts for each word.

Technology

Big Data & Cloud

The Hadoop Ecosystem



Pig

- Data-flow language; simplifies MapReduce programming

Hive and HiveQL

- SQL-like language supports defining tables and issuing SQL-like queries

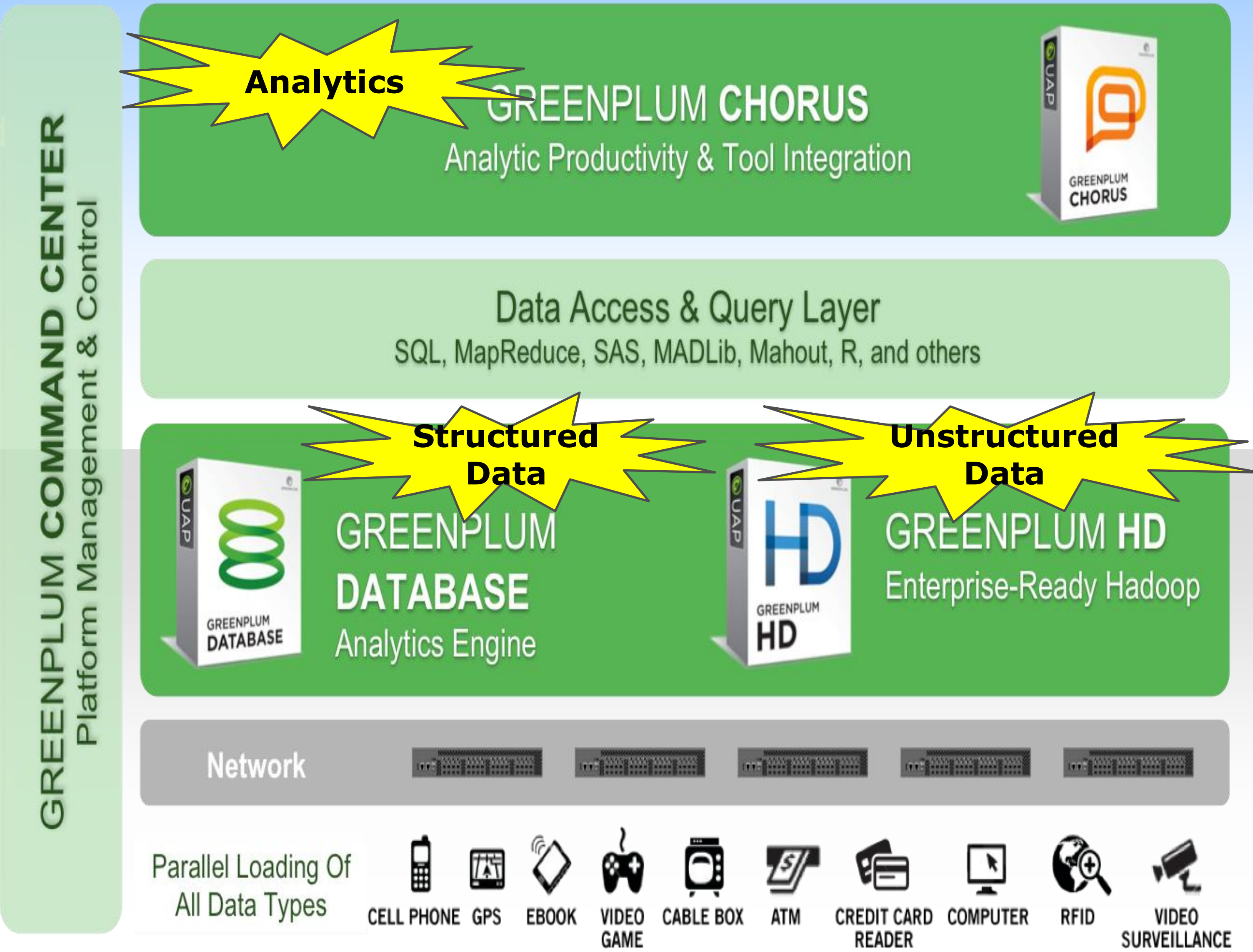
HBase

- BigTable: millions of rows, millions of columns
- Provides single record access and updates

All of the above leverage Hadoop's MapReduce framework and HDFS

Tools

Big Data & Clouc



Analytics

GREENPLUM CHORUS

Analytic Productivity & Tool Integration

Data Access & Query Layer

SQL, MapReduce, SAS, MADLib, Mahout, R, and others

Structured Data

Unstructured Data

GREENPLUM DATABASE

Analytics Engine

GREENPLUM HD

Enterprise-Ready Hadoop

Network

GREENPLUM COMMAND CENTER

Platform Management & Control

Parallel Loading Of All Data Types

Conclusion

Big Data challenges

1) Need for highly-scalable systems
2) Need for highly-available systems
3) Demand huge hardware investment

Cloud benefits

1) Provide highly flexible and scalable systems
2) Provide higher availability for applications
3) Reduce costs

Big Data + Cloud + Technology =

Affordable Cost + Better Analytics + Competitive Advantage

Explore Bigger Possibilities

EMC²

Conclusion

Think Big

Think Data

Think Big Data

EMC2

References

- Forrester Reports on Big Data and Cloud Computing
- Gartner Reports on Big data and Cloud Computing
- EMC Internal Big Data and Cloud Computing Initiatives and Education
- IDC Reports and analysis on big data and cloud
- National Institute of Standards and Technology (NIST), Information Technology Laboratory]
- Patricia Florissi, CTO, EMC Sales

EMC²

Q & A

EMC^2®