Big Data in the Enterprise. When to Use What?

Jesus Rodriguez, Tellago, KidoZen, Inc

tellago

Agenda

- Big Data principles
- The Hadoop ecosystem
- Other big data technologies

tellago

About Me

- Co-Founder Tellago, Inc
- Co-Founder KidoZen, Inc
- Microsoft MVP
- Architect Advisor
- Investor
- Speaker, Author
- http://jrodthoughts.com
- http://weblogs.asp.net/gsusx
- http://kidozen.com

tellago

About Tellago

- Application development firm focused on big enterprise trends (launched 2008)
  - Enterprise mobility, cloud computing, augmented reality, modern BI & big data
- Advisor to software companies such as Microsoft or Oracle
- American Business Awards(2011) “Best Overall Company of the Year < 100”
- American Business Awards(2012) Silver: “Best Computer Services Company of the Year < 100”, Silver: Best Computer Services Executive of the Year
- Inc 500 (114) & other industry awards

tellago

Some Housekeeping Rules

- Tellago Technology Updates focused on modern enterprise software trends
- Real world stories
- No sales pitch
- Leverage GTW to ask questions

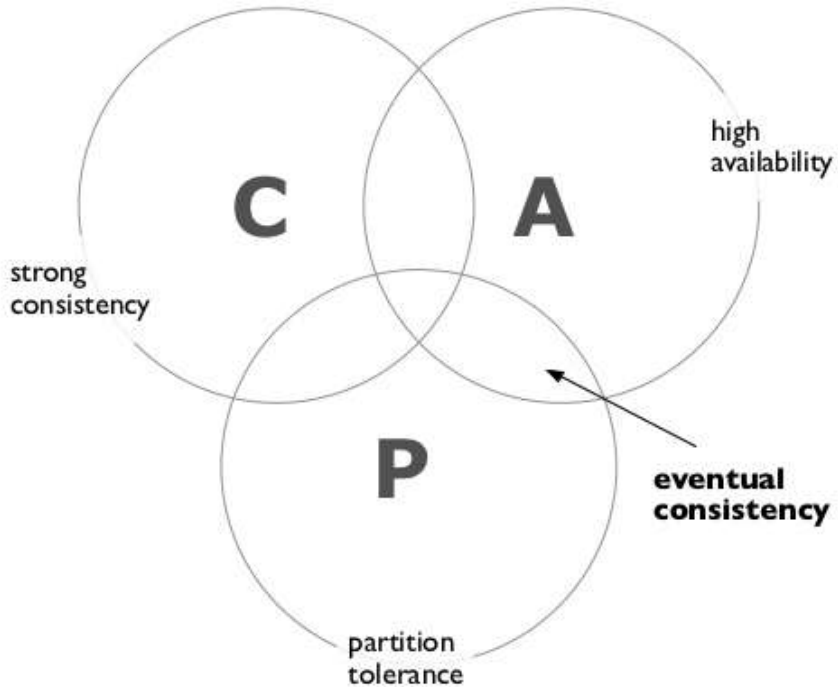
tellago

We Love Data!

Where all Started?

tellago

CAP Theorem



"You can have at most two of these properties for any shared-data system… the choice of which feature to discard determines the nature of your system." – Eric Brewer, 2000 (Inktomi)

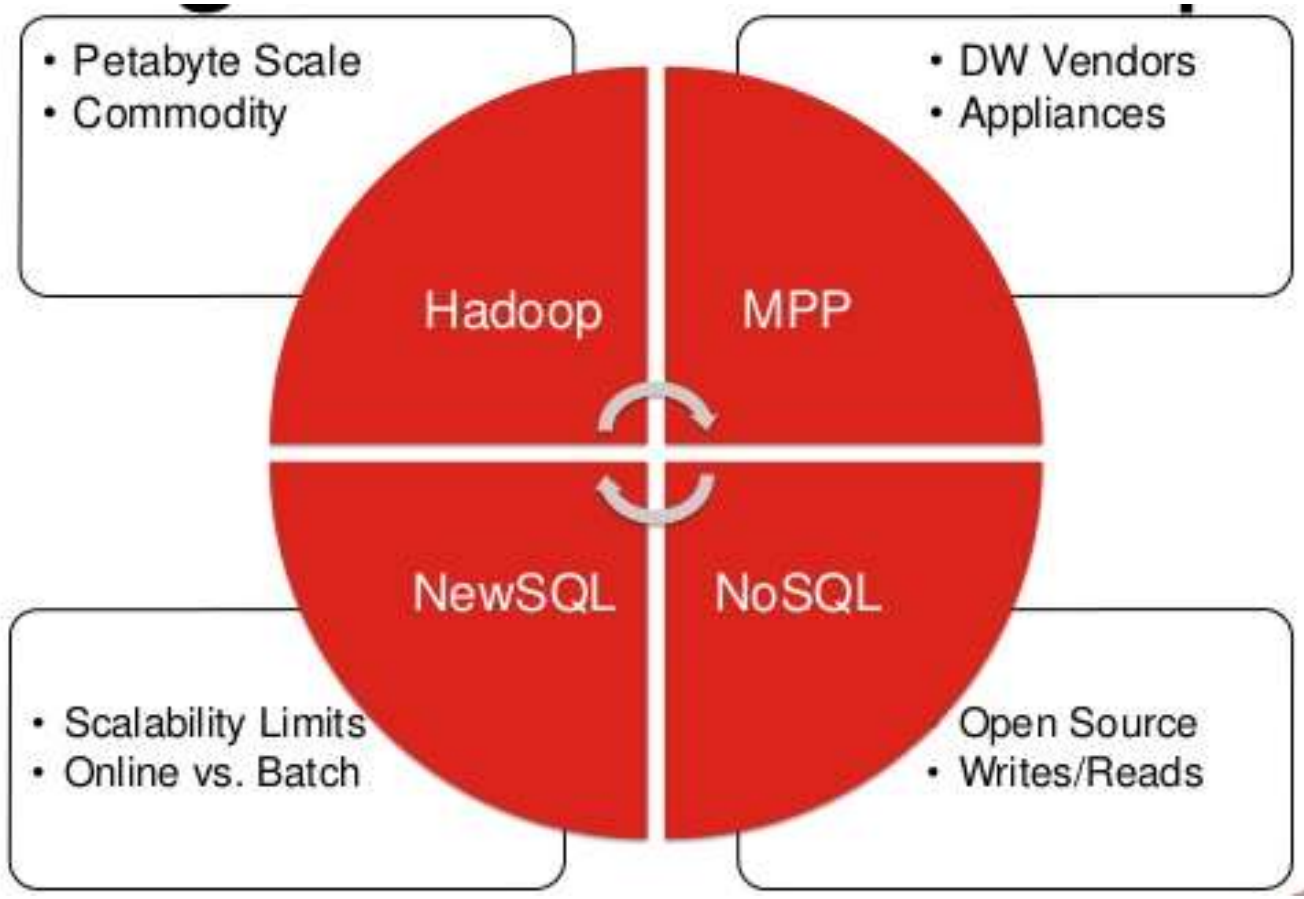
- revenue transactions in ecommerce typically require **strong consistency** and **partition tolerance**
- most analytics jobs for business use cases generally require **availability** and **eventual consistency**, but tend to not tolerate highly partitioned data
- ETL becomes an Achilles heal for "agile":
  - agile/experiment-driven/scale-out, which leads to…
  - provably-hard-to-detect metadata drift, leading to…
  - high-risk technical debt

tellago

Big Data Opportunity

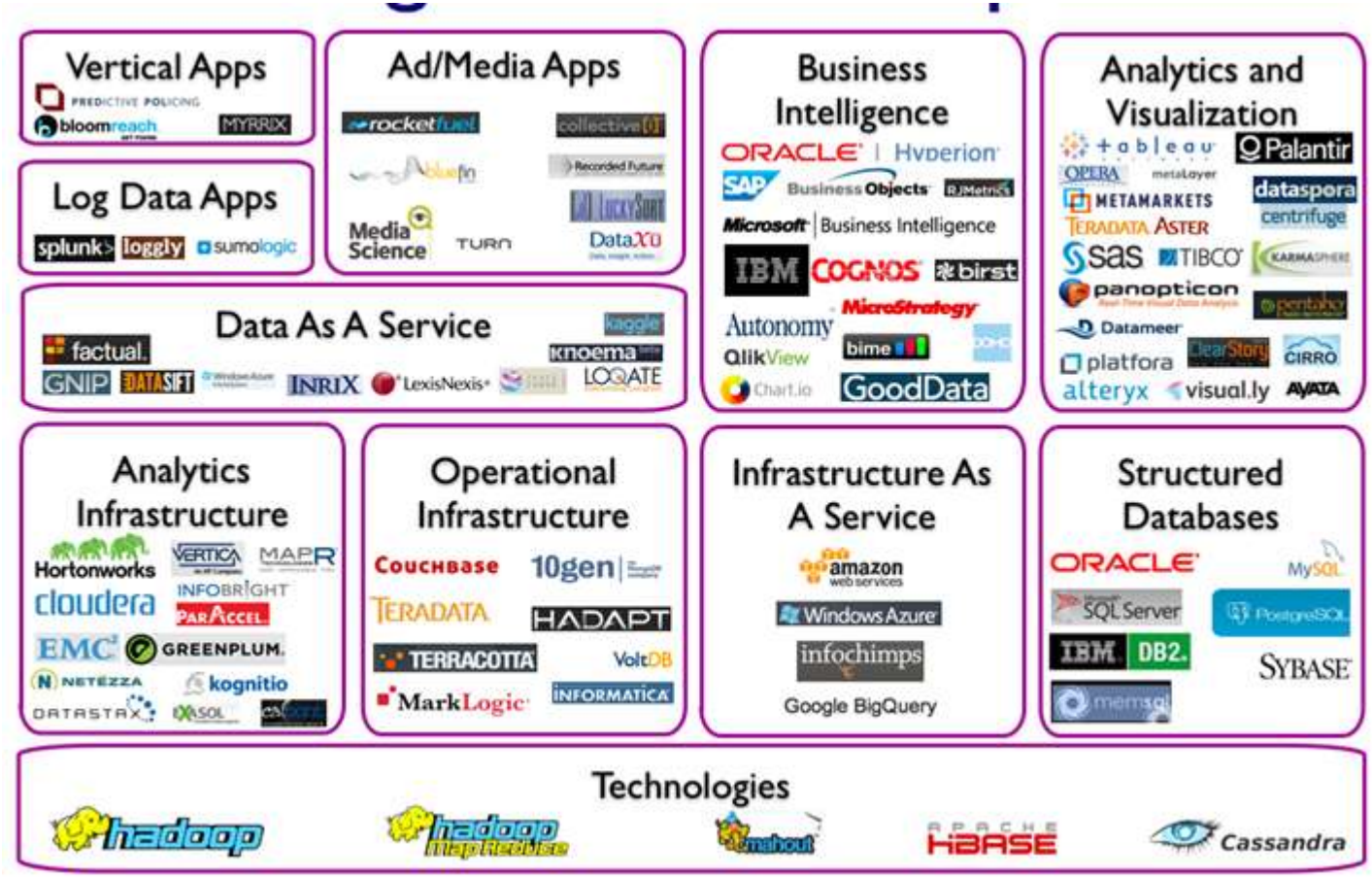
tellago

The Landscape



tellago

Or a Bit More Crowded



tellago

Or Worse

tellago

Hadoop Led the Way

tellago

Hadoop Design Principles

- System Shall Manage and Heal Itself
- Performance Shall Scale Linearly
- Compute Shall Move to Data
- Simple Core, Modular and Extensible

The Solution: HDFS + Map Reduce

Distributed File System

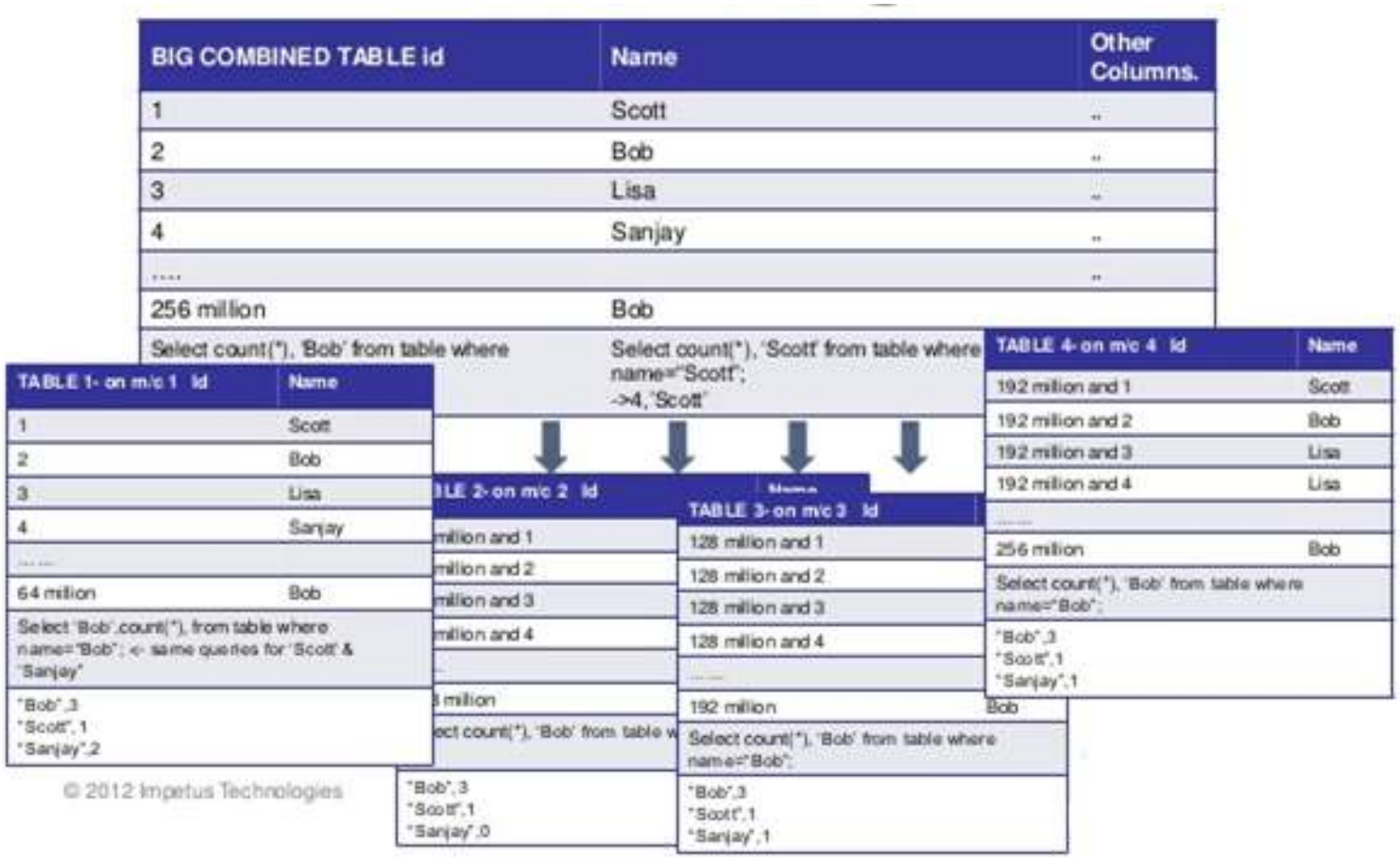
- Petabyte Scale
- Thousands of Commodity Servers
- High Availability
- Highly Fault Tolerant

Distributed Processing System

- Simple easy to code Algorithm
- Code once Run on PBs
- High Fault Tolerance
- Data Locality

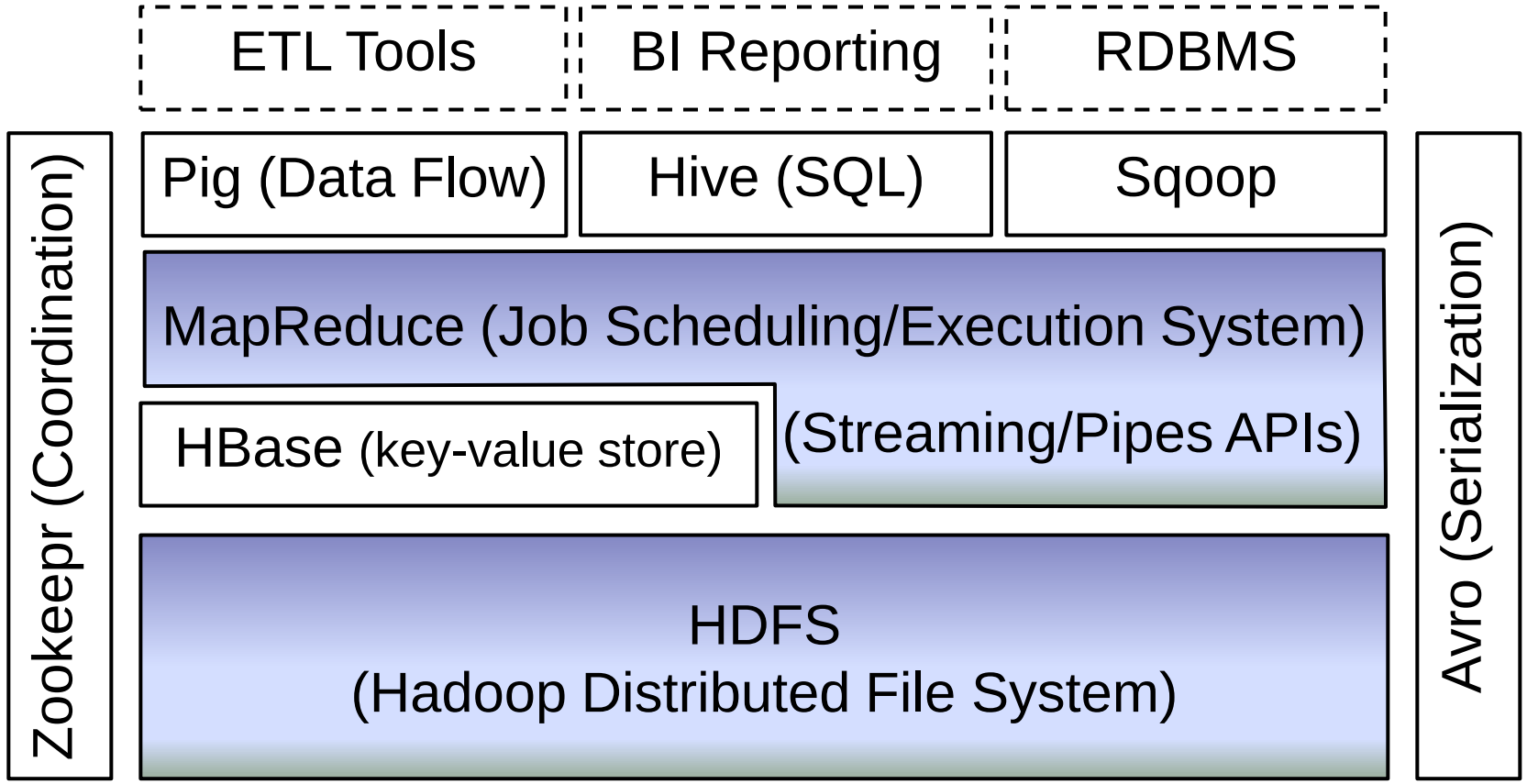
tellago

Mapping



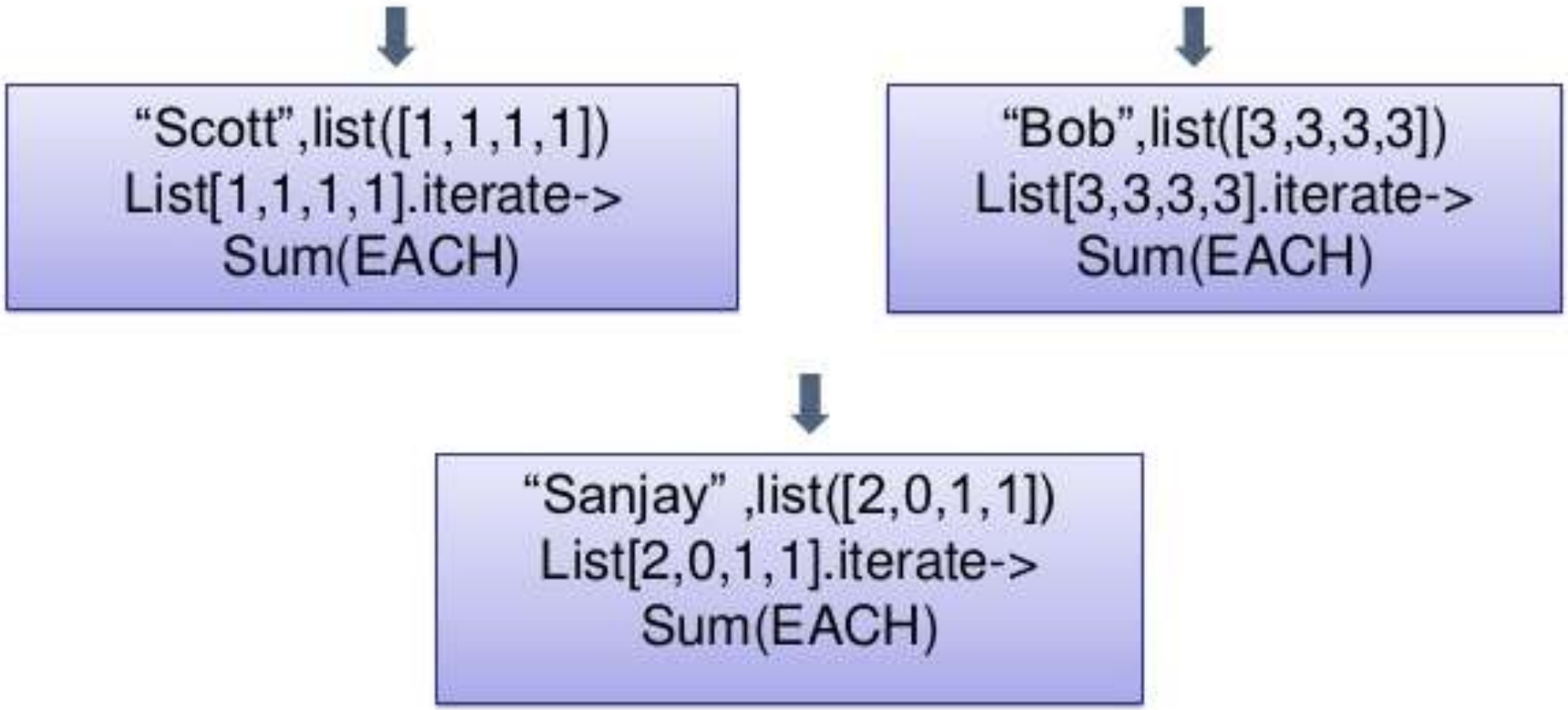
tellago

Hadoop Ecosystem

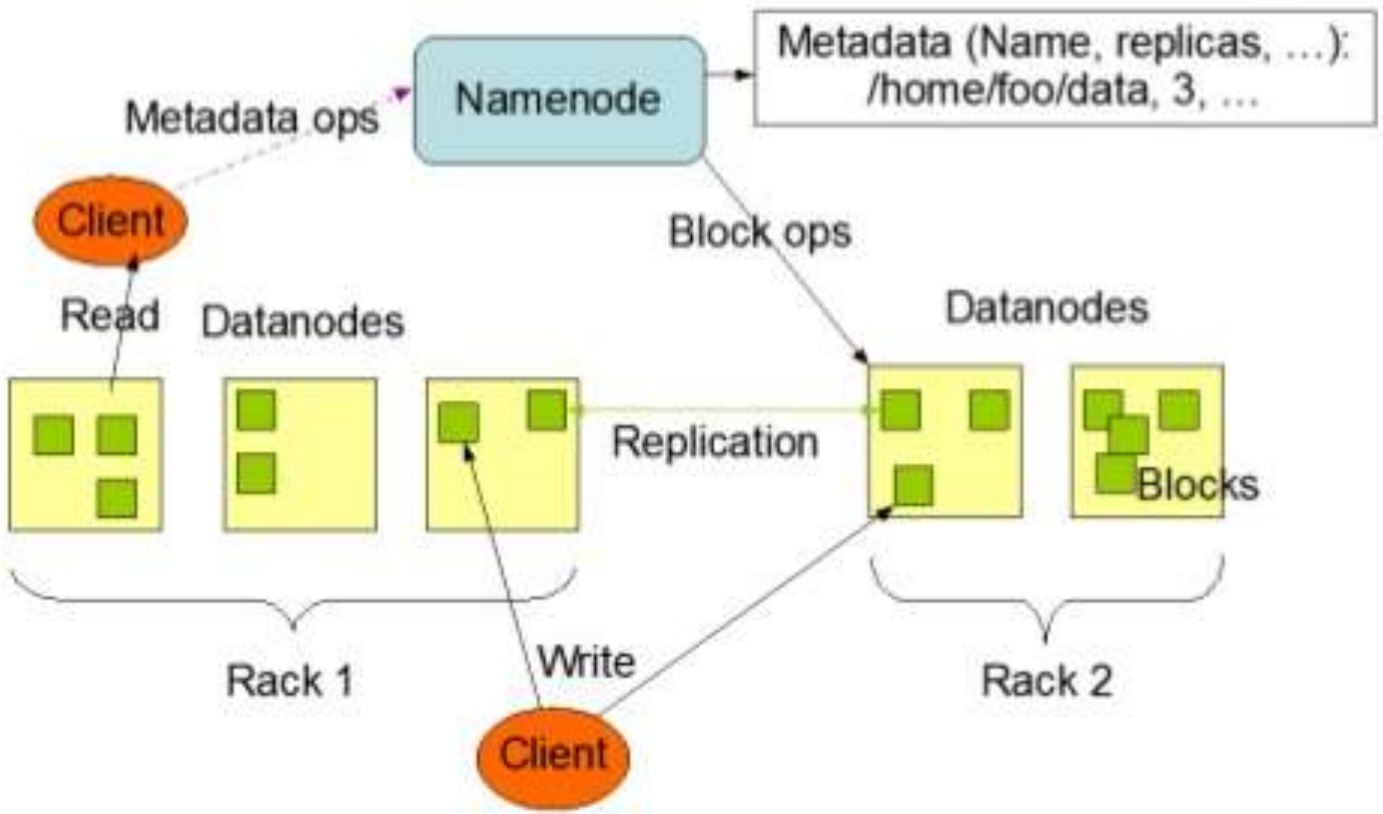


tellago

Reducing


"Scott",list([1,1,1,1])
List[1,1,1,1].iterate->
Sum(EACH)
"Bob",list([3,3,3,3])
List[3,3,3,3].iterate->
Sum(EACH)
"Sanjay" ,list([2,0,1,1])
List[2,0,1,1].iterate->
Sum(EACH)

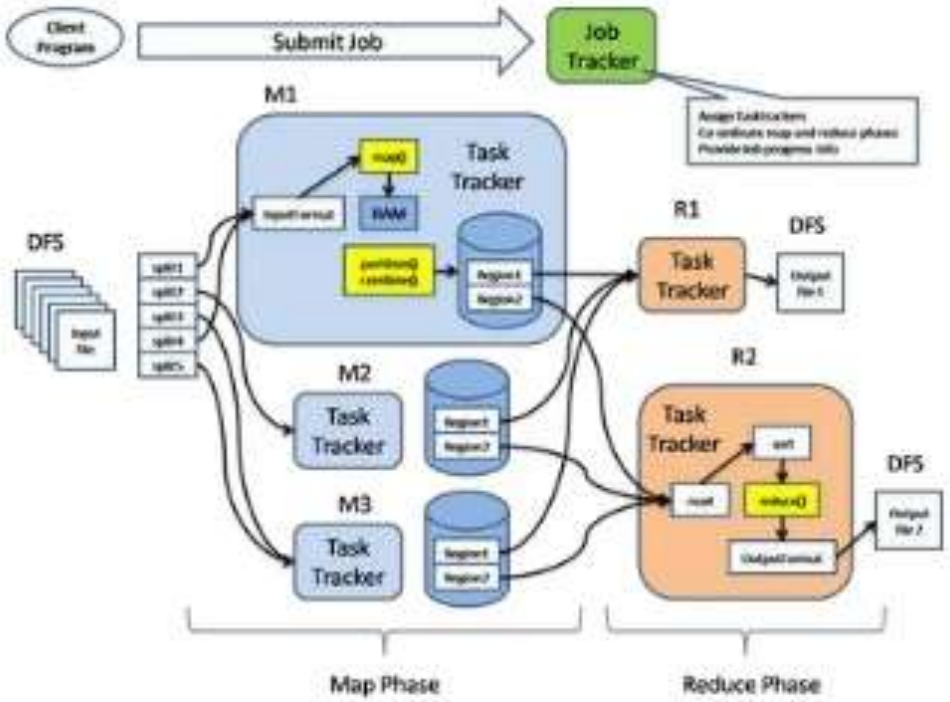
HDFS



tellago

Map-Reduce

map (k1,v1) → list(k2,v2)
reduce (k2,list(v2)) → list(v2)



tellago

Relational vs. Hadoop

Relational	VS.	Hadoop
Required on write	schema	Required on read
Reads are fast	speed	Writes are fast
Standards and structured	governance	Loosely structured
Limited, no data processing	processing	Processing coupled with data
Structured	data types	Multi and unstructured
Interactive OLAP Analytics Complex ACID Transactions Operational Data Store	best fit use	Data Discovery Processing unstructured data Massive Storage/Processing

tellago

The Hadoop Ecosystem

tellago

WebHDFS

Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

WebHDFS

- REST API that supports the complete FileSystem interface for HDFS.
- Move data in and out and delete from HDFS
- Perform file and directory functions

- webhdfs://<HOST>:<HTTP PORT>/PATH

- Standard and included in version 1.0 of Hadoop

tellago

Sqoop

Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Sqoop

- Sqoop is a set of tools that allow non-Hadoop data stores to interact with traditional relational databases and data warehouses.
- A series of connectors have been created to support explicit sources such as Oracle & Teradata
- It moves data in and out of Hadoop
- SQ-OOP: SQL to Hadoop

tellago

Flume

Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Flume

- Distributed service for efficiently collecting, aggregating, and moving streams of log data into HDFS
- Streaming capability with many failover and recovery mechanisms

- Often used to move web log files directly into Hadoop

tellago

HBase

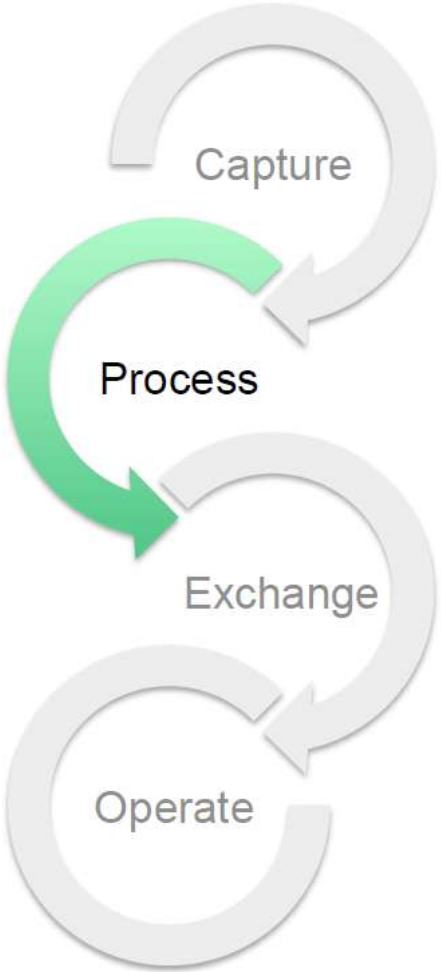
Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache HBase

- HBase is a non-relational database. It is columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes.

tellago

Pig



Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Pig

- Apache Pig allows you to write complex map reduce transformations using a simple scripting language. Pig latin (the language) defines a set of transformations on a data set such as aggregate, join and sort among others. Pig Latin is sometimes extended using UDF (User Defined Functions), which the user can write in Java and then call directly from the language.

tellago

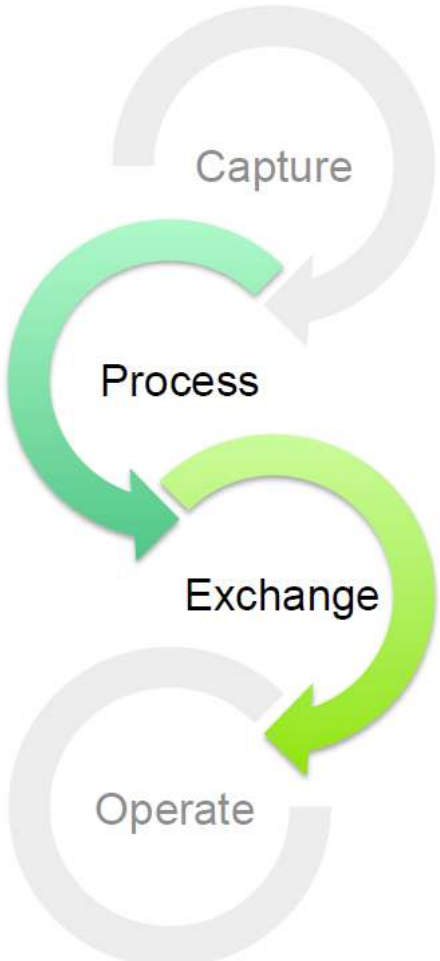
HCatalog

Apache HCatalog

- HCatalog is a metadata management service for Apache Hadoop. It opens up the platform and allows interoperability across data processing tools such as Pig, Map Reduce and Hive. It also provides a table abstraction so that users need not be concerned with where or how their data is stored.

tellago

Hive



Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Hive

- Apache Hive is a data warehouse infrastructure built on top of Hadoop (originally by Facebook) for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL (HQL).

tellago

Ambari

Apache Ambari

- Ambari is a monitoring, administration and lifecycle management project for Apache Hadoop clusters
- It provides a mechanism to provisions nodes

- Operationalizes Hadoop.

tellago

Oozie

Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Oozie

- Oozie coordinates jobs written in multiple languages such as Map Reduce, Pig and Hive. It is a workflow system that links these jobs and allows specification of order and dependencies between them.

tellago

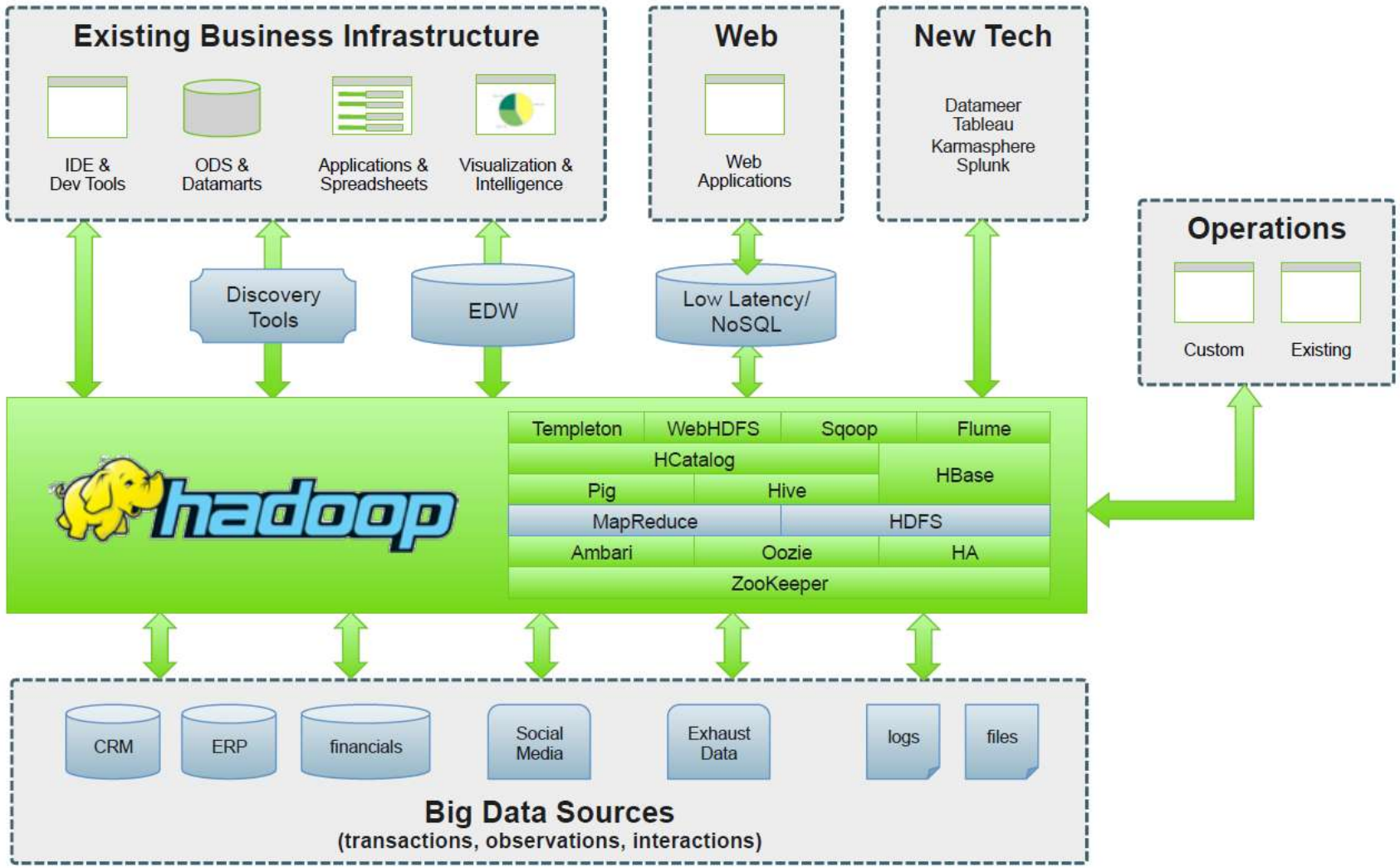
Zookeeper

Apache ZooKeeper

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

tellago

Hadoop Enterprise Architecture



tellago

Hadoop is not a silver bullet...

tellago

Some Challenges

- **Hadoop doesn't power big data *applications***
  - Not a transactional datastore. Slosh back and forth via ETL
- **Processing latency**
  - Non-incremental, must re-slurp entire dataset every pass
- **Ad-Hoc queries**
  - Bare metal interface, data import
- **Graphs**
  - Only a handful of graph problems amenable to MR

tellago

Beyond Hadoop

- **Percolator(**incremental processing)

http://research.google.com/pubs/pub36726.html

- **Dremel(**ad-hoc analysis queries)

http://research.google.com/pubs/pub36632.html

- **Pregel (**Big graphs)

http://dl.acm.org/citation.cfm?id=1807184

tellago

Important Big Data Technologies in the Enterprise

tellago

Real Time Analytics

tellago

Real Time Analytics

- Storm
- Hstreaming
- StreamBase
- IBM Streams
- Microsoft StreamInsight

tellago

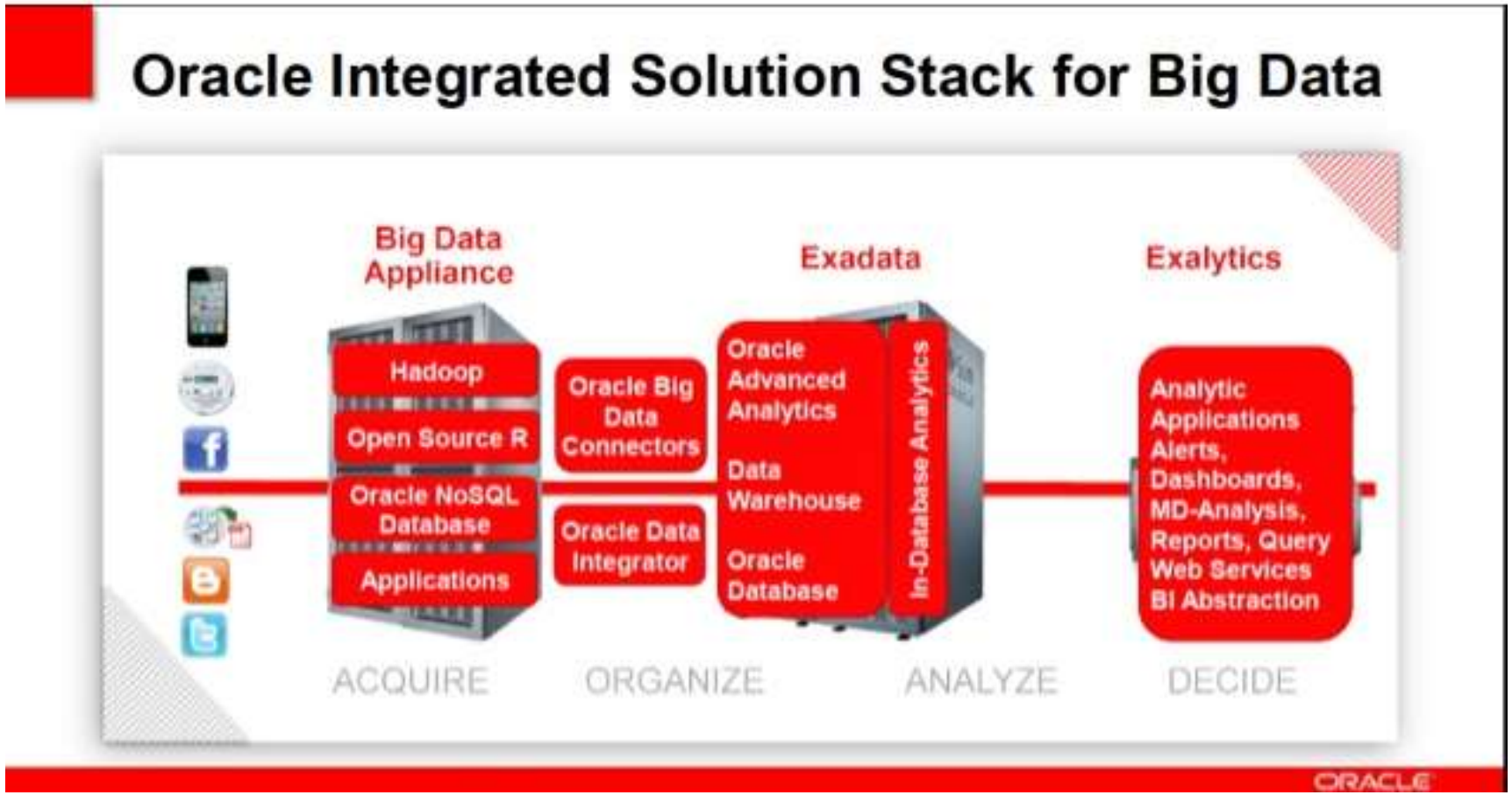
MPP: Massively Parallel Processing

tellago

MPP Columnar Stores

- Oracle Exadata
- IBM Netezza
- Teradata
- EMC Greenplum
- HP Vertica
- ParAccel
- Microsoft SQL Server PDW

Oracle & Big Data



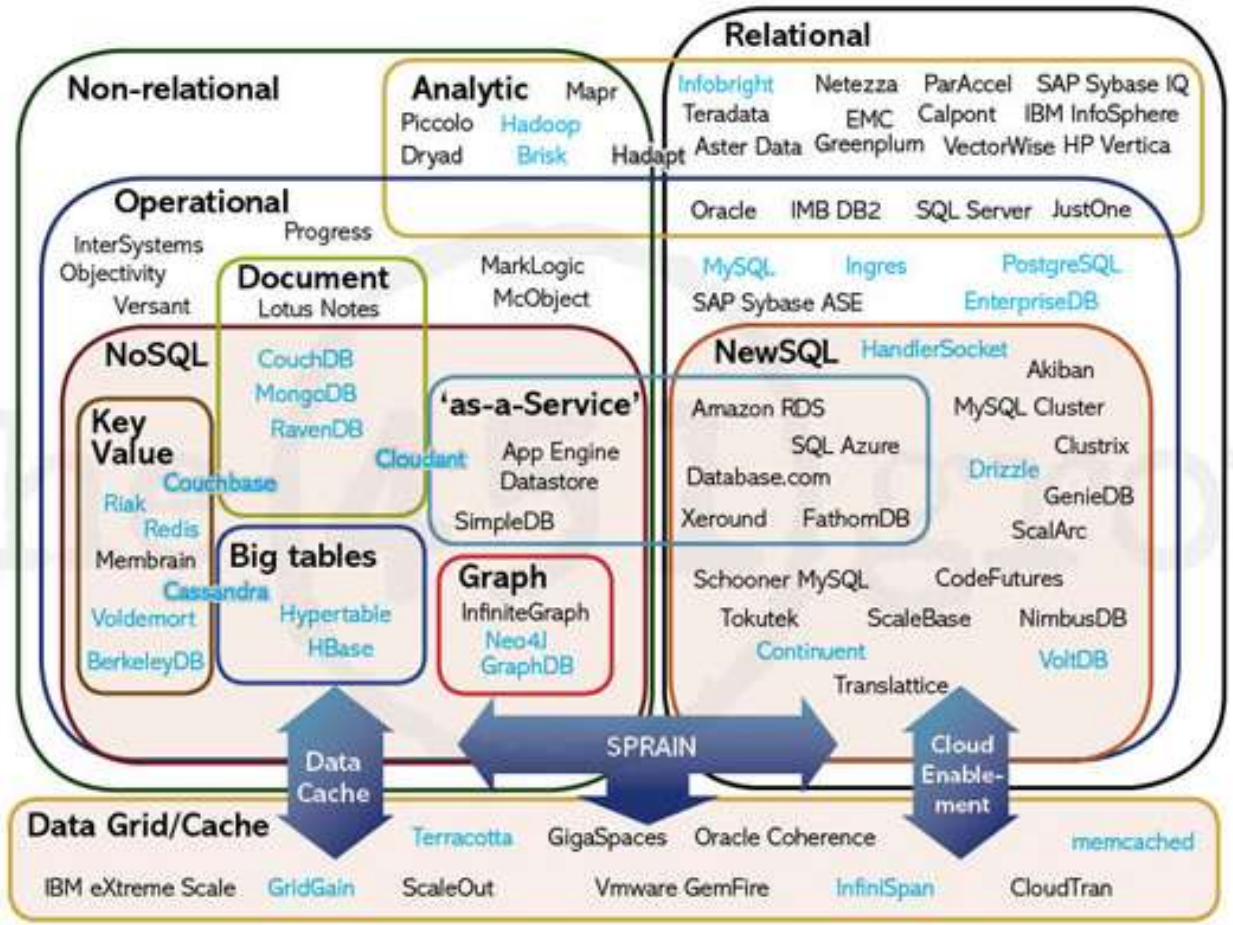
tellago

Microsoft & Big Data

tellago

NoSQL DBs

NoSQL DBs



tellago

NewSQL DBs

New SQL / Cloud DB

- VoltDB
- NimbusDB
- SimpleDB
- NuoDB
- Clustrix
- Totutek

tellago

Traditional BI Suites

tellago

New SQL / Cloud DB

- Hadoop Support In:
  - Microsoft SSIS
  - Informatica Datastage
  - Talend
  - Pentaho
  - Microstrategy , SaaS
  - Tableau, Qlikview

tellago

Big Data & Cloud

tellago

Big Data & Cloud

- Hadoop distributions (AWS, Microsoft HDInsight, Cloud Foundry)
- Data marketplaces (Factual, Infochimps)
- Data visualization (WibiData)
- NOSQL as a Service (MongoHQ)

If you are interested on evaluating Big Data in your organization

tellago

Tellago Big Data Strategy Session

- 1 day strategy session
- Start with a real world scenario
- Explore various big data technology vendors
- Present a potential technology roadmap
- Free

- Emails us at info@tellago.com

tellago

Summary

- The big data ecosystem is super crowded
- Hadoop distributions are leading the way in the enterprise
- Complementary technologies include:
  - NOSQL
  - New SQL
  - MPP
  - Data Visualization

tellago

Thanks

jesus.rodriguez@tellago.com
http://www.tellagostudios.com
http://jrodthoughts.com
http://twitter.com/#!/jrodthoughts
http://weblogs.asp.net/gsusx

tellago