The age of Big Data

Big Data for Oracle Database Professionals

Oracle OpenWorld 2017 #OOW17

SessionID: SUN5698

dataREDDY

Tom S. Reddy

tom.reddy@datareddy.com

About the Speaker

- COLLABORATE & OpenWorld Speaker
- IOUG COLLABORATE Conference Committee…2018

- Oracle Certified Database Administrator (OCP)
- aws Certified Solutions Architect…in progress
- Hadoop & Spark…in progress

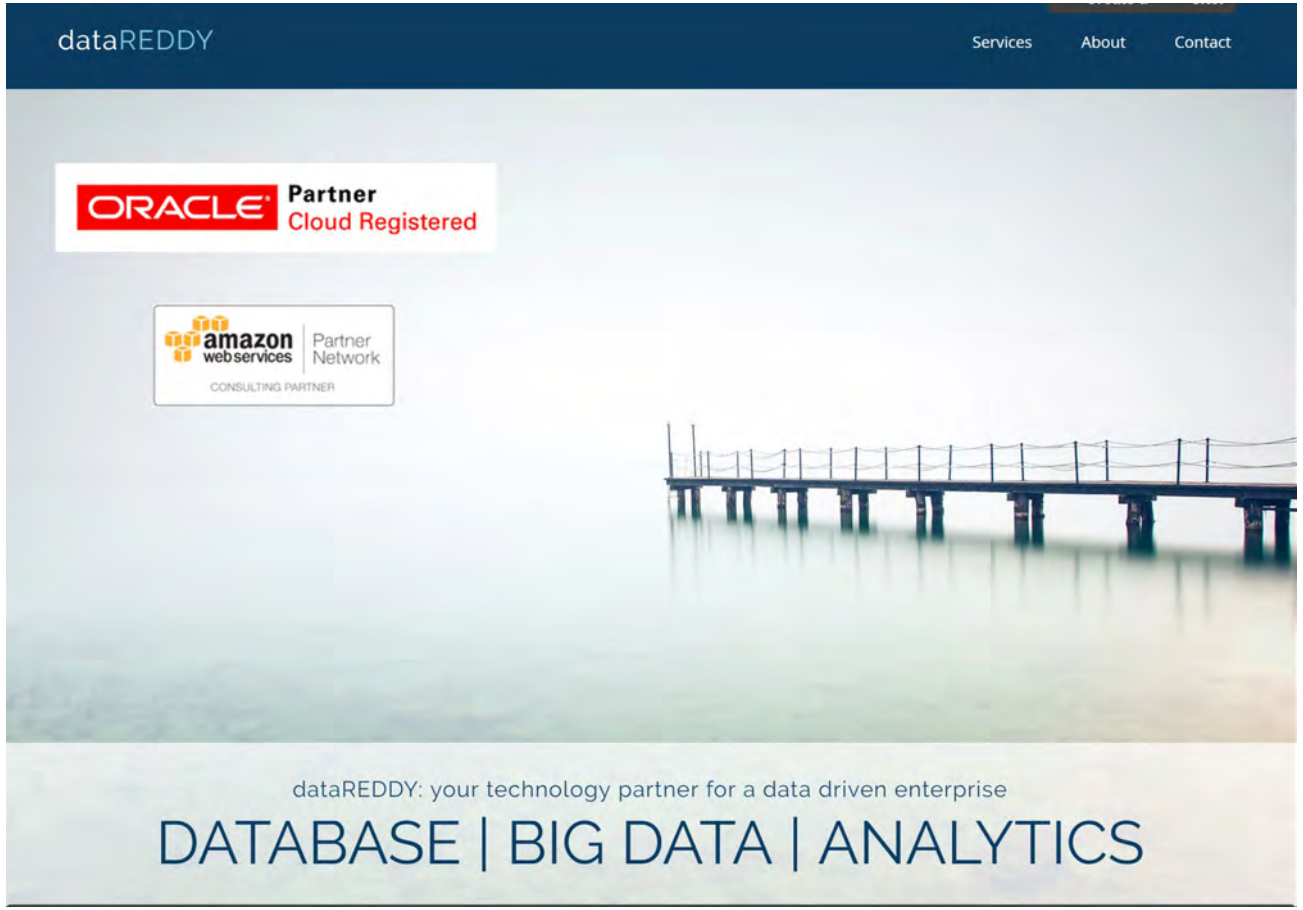
- Email: tom.reddy@datareddy.com
- LinkedIn: https://www.linkedin.com/in/tomreddy/
- Twitter: @tomreddydba

About the Company

- data platforms

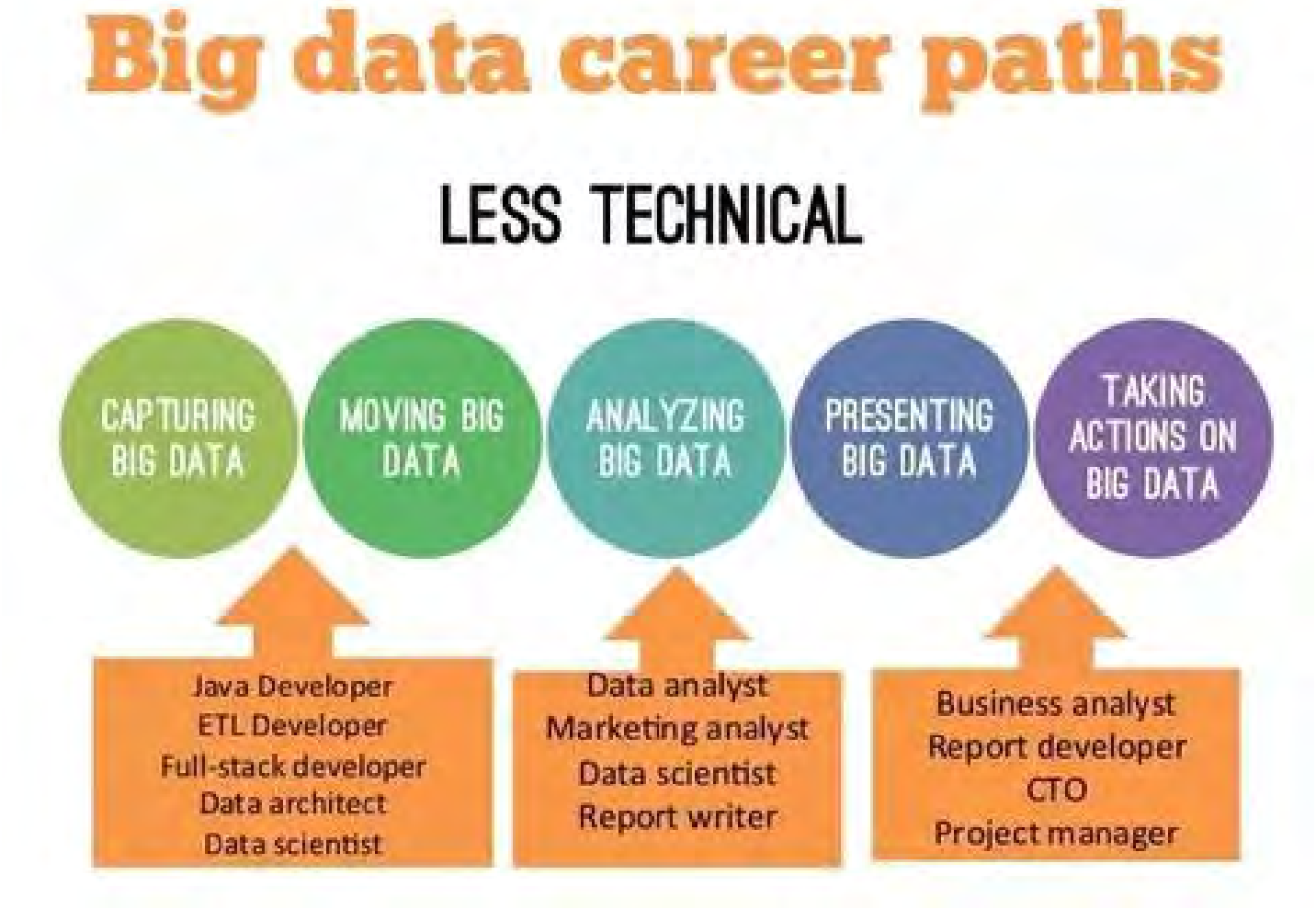
- Oracle
- Hadoop
- Cloud
- Analytics

- www.datareddy.com



Survey

- Primary Focus
  - Oracle
    - DBA
    - Developers
    - Manager/Others
- Hadoop
- Cloud
- Data Engineering
- Data Scientists
- AI/ML



Best of Breed

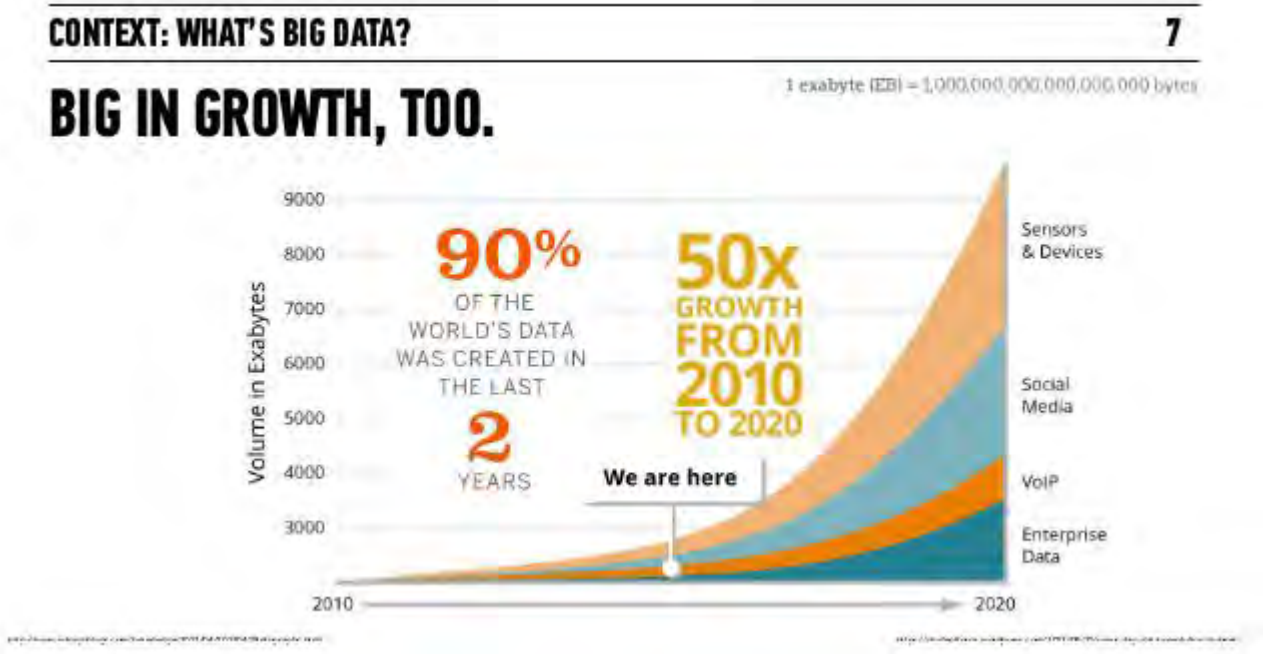
- Polyglot Persistence
  - Pick the right storage & engine for the right use case
  - Hybrid: Cloud vs On-Prem

- The purpose of data platforms remains the same:
  - Store data
  - Retrieve data
  - Analyze/Process data
  - How efficiently we do this depends on the platform!

What is Big Data?

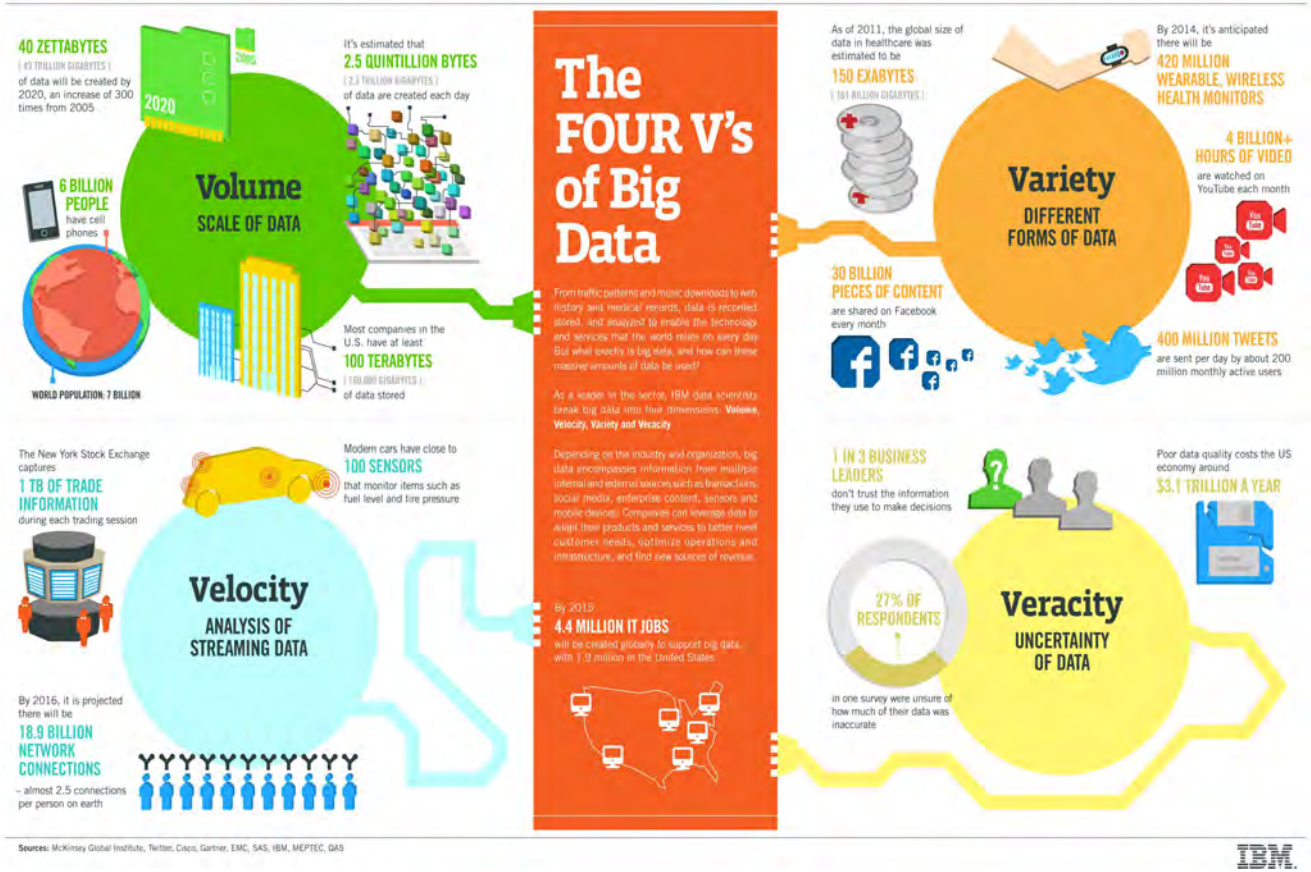
- Large Data sets that are difficult to manage with traditional database tools

- Petabytes of data
- IoT
- Streaming data
- BLOB's: Images, Videos
- Unstructured/Semi-structured



Big Data – 4 V's

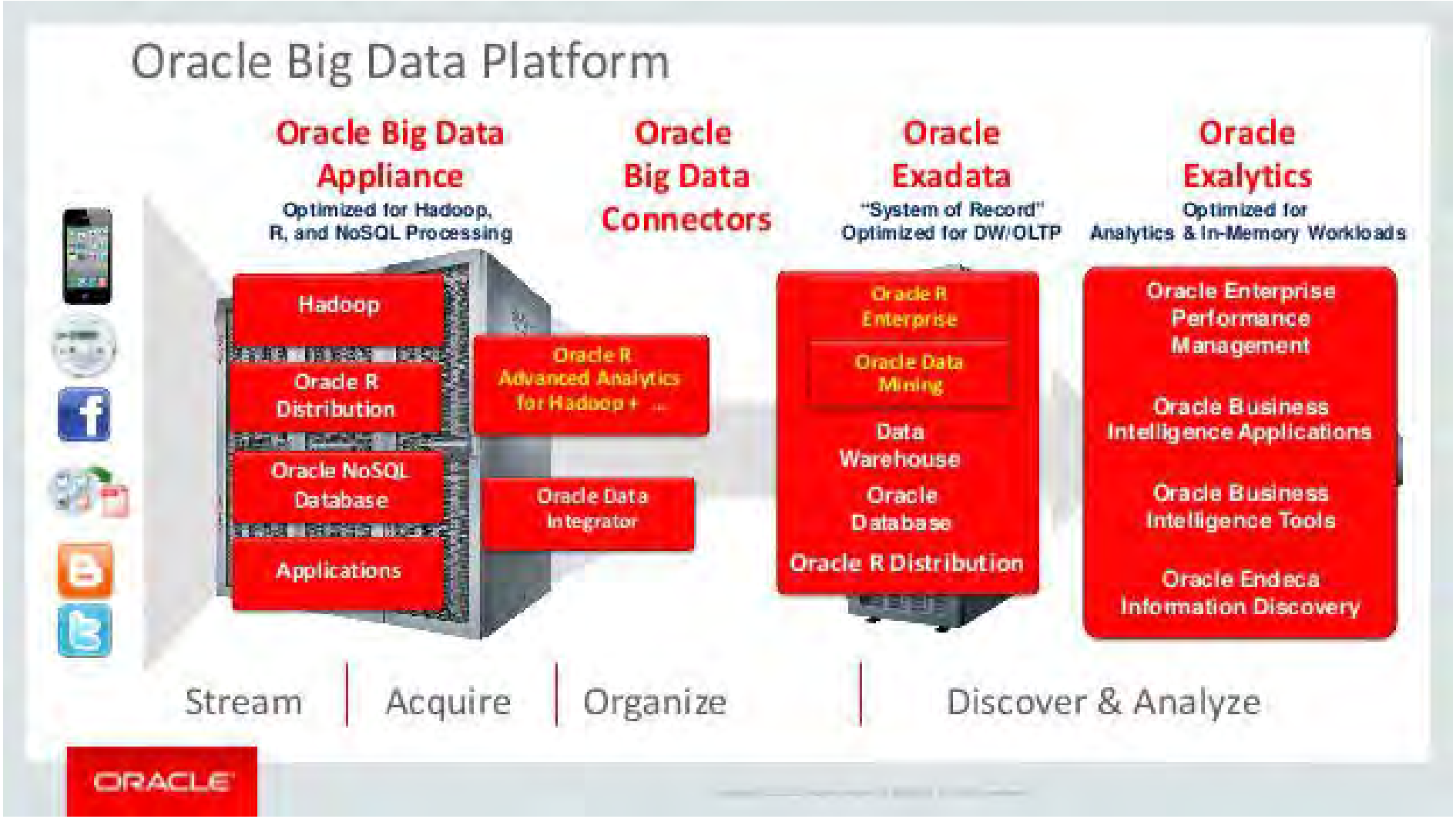
- Volume
  - Scale of Data
- Velocity
  - Analysis of Streaming Data
- Variety
  - Different forms of Data
- Veracity
  - Uncertainty of Data



Why Big Data?

- Scalability
- Distributed processing
- Lower costs – typically,
- Lower processing times
- Real-Time Analytics
- Machine Learning
- AI

Oracle Big Data Platform



Oracle Big Data Cloud Service

- On-Demand
- Oracle hosted Big Data Machine/Appliance
  - Hadoop
  - Spark
- Fast Big Data Connectors/Integration
  - Oracle Data Integrator
  - Oracle SQL Connector
  - Oracle Loader for Hadoop
- Big Data Spatial & Graph
- Big Data SQL

Big Data Products - Oracle

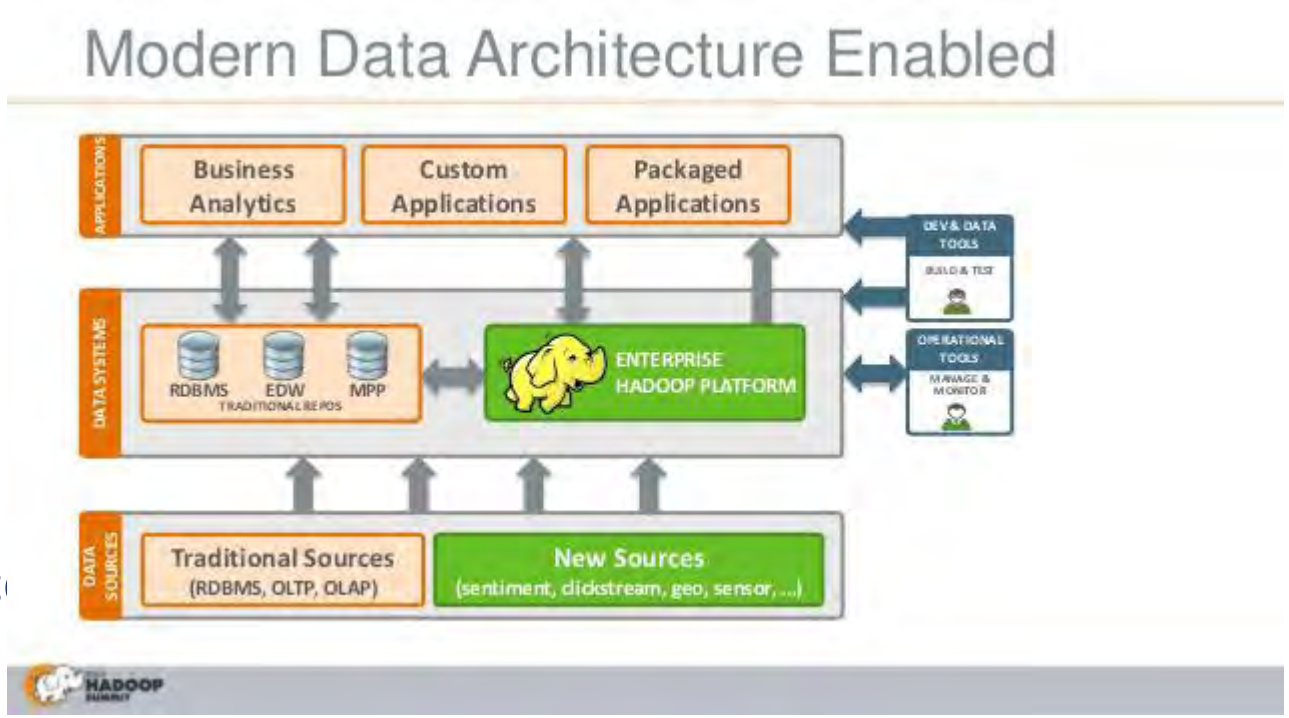
- Aggregate
  - Oracle Big Data Preparation Cloud Service
  - Oracle IoT Cloud Service
  - Oracle Golden Gate Cloud Service
- Manage
  - Oracle Big Data Cloud Service
  - Oracle Big Data Cloud Machine
- Experiment
  - Big Data Discovery Cloud Service
  - Oracle R for Hadoop

Big Data Platforms - Overview

- Apache – Open Source
  - Hadoop
    - Distributed storage and processing of big data sets using MapReduce
  - Spark
    - Distributed cluster-computed/data-processing framework
- NoSQL
  - Non-relational data storage
- Graph
  - Manage highly connected data & queries
- Cloud
  - S3
    - Cloud-based low-cost object storage

Big Data Platforms…Cont'd

- Apache Hadoop
  - Cloudera
  - MapR
  - HortonWorks
- Apache Spark
  - DataBricks
  - Distributed General Data Proc
- aws/s3



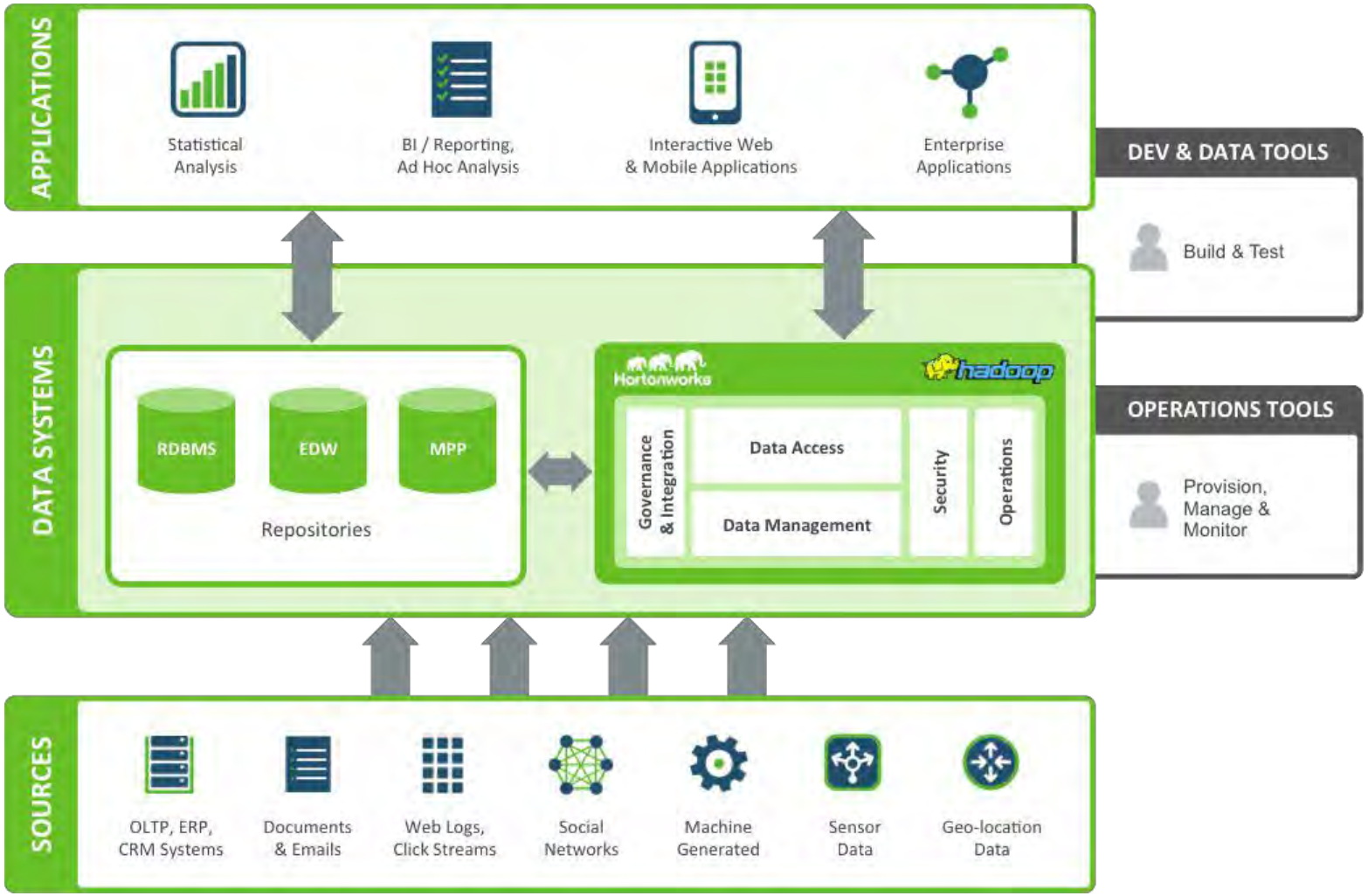
Big Data Platforms

- No SQL
  - Document
    - Mongo
    - Couchbase
  - Graph
    - Neo4j
    - Giraph
  - Key-Value Stores
    - Riak
    - Berkeley DB
  - Wide-Column Stores
    - Cassandra
    - HBase

Amazon – aws Big Data

- Cloud Scale services include:
- EMR
  - Hadoop & Spark
- S3
  - Big Data Object Storage
- DynamoDB
  - Managed NoSQL
- Graph Databases
- Lex

Hadoop in the enterprise

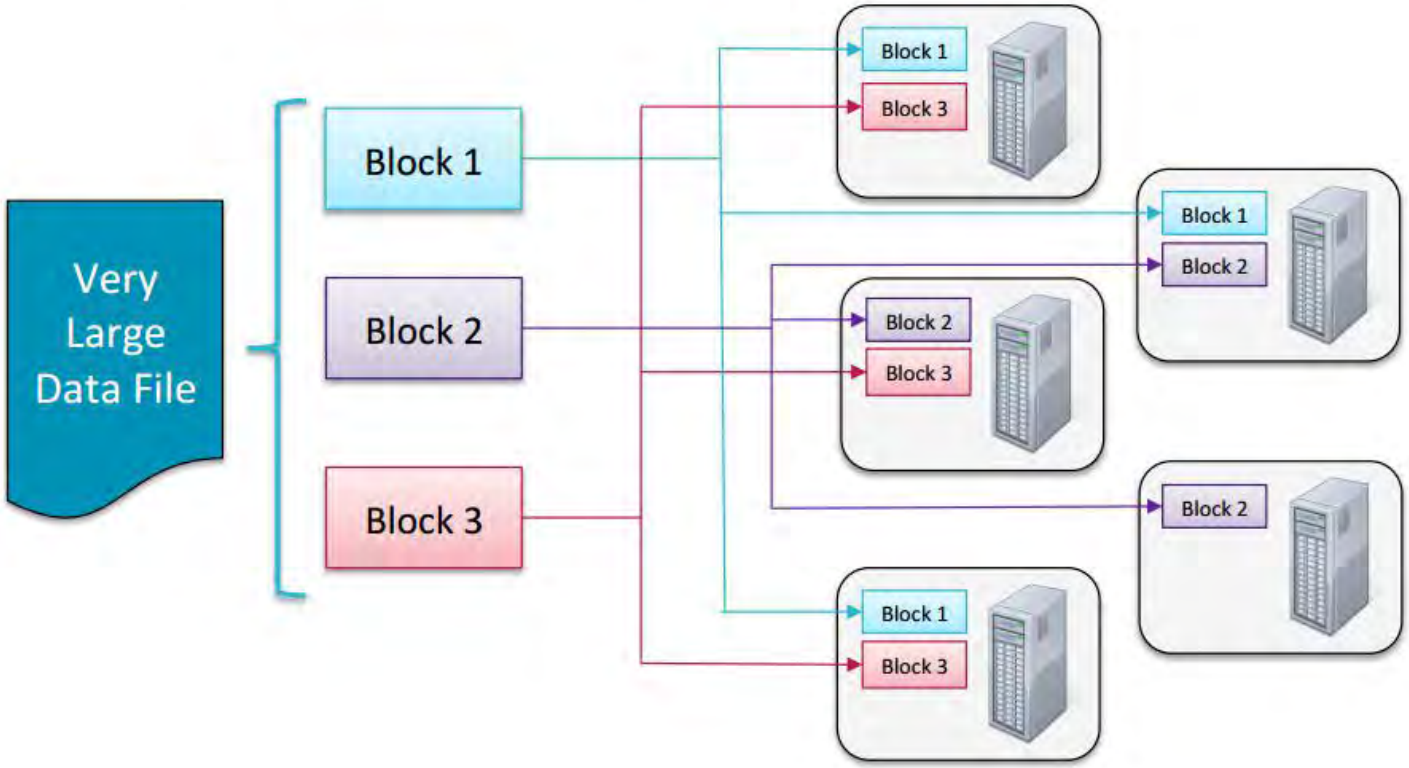


Big Data Platforms – Hadoop

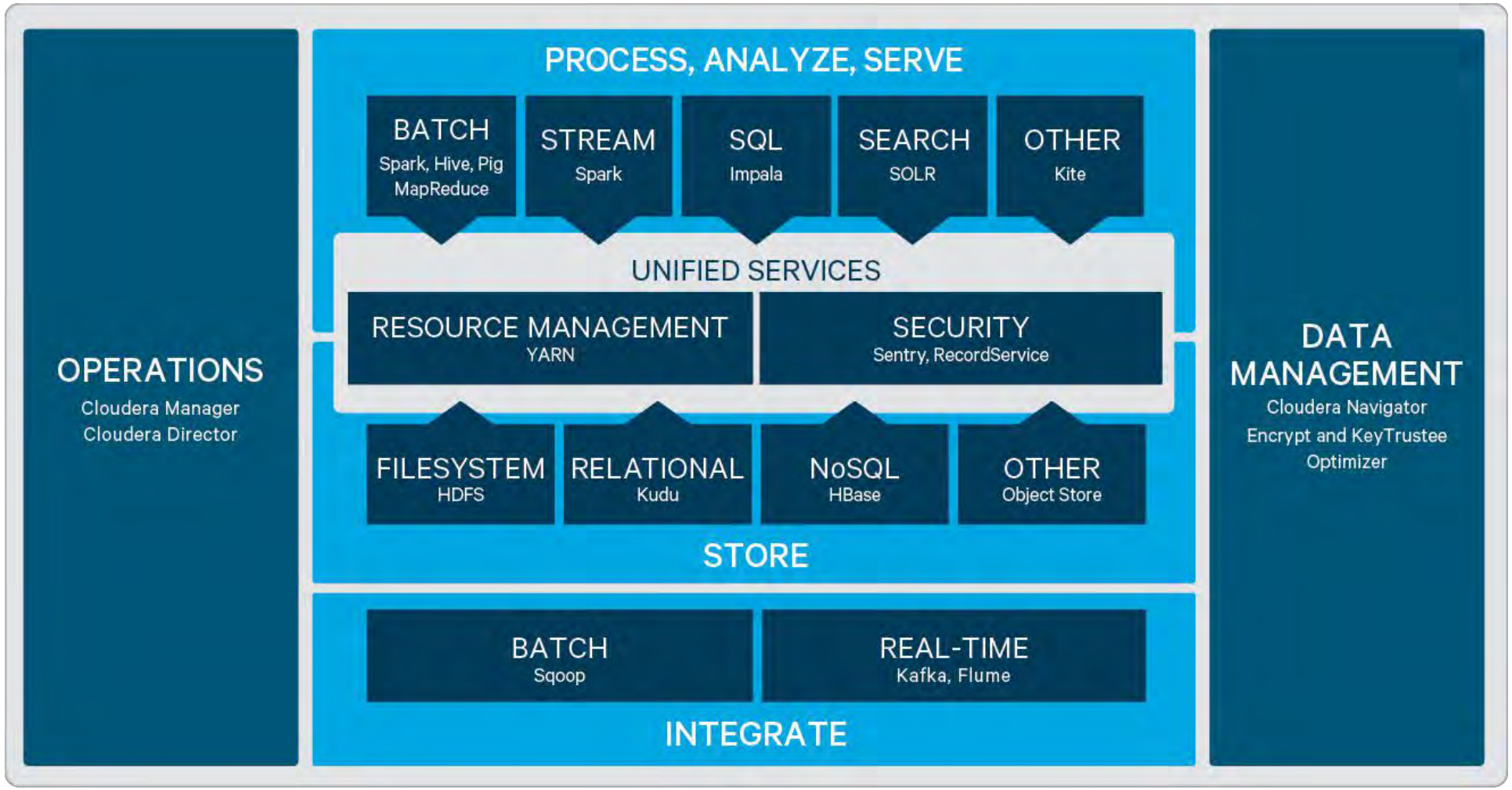
- Distributed File System
- Specialized to facilitate storage & processing of large volumes of data
- Unstructured & Streaming data
- Massively Scalable
- Highly Available
- Active Ecosystem
- Enterprise Grade

Hadoop Storage

- **Data files are split into blocks and distributed to data nodes**
- **Each block is replicated on multiple nodes (default 3x)**



Cloudera Ecosystem



Cloudera Ecosystem Project descriptions

Project	What does it do?
Spark	In-memory execution framework
HBase	NoSQL database built on HDFS
Hive	SQL processing engine designed for batch workloads
Impala	SQL query engine designed for BI workloads
Parquet	Very efficient columnar data storage format
Sqoop	Data movement to/from RDBMSs
Flume, Kafka	Streaming data ingestion
Solr	Enables users to find the data they need
Hue	Web-based user interface for Hadoop
Oozie	Workflow scheduler used to manage jobs
Sentry	Authorization tool, providing security for Hadoop

MapR Ecosystem

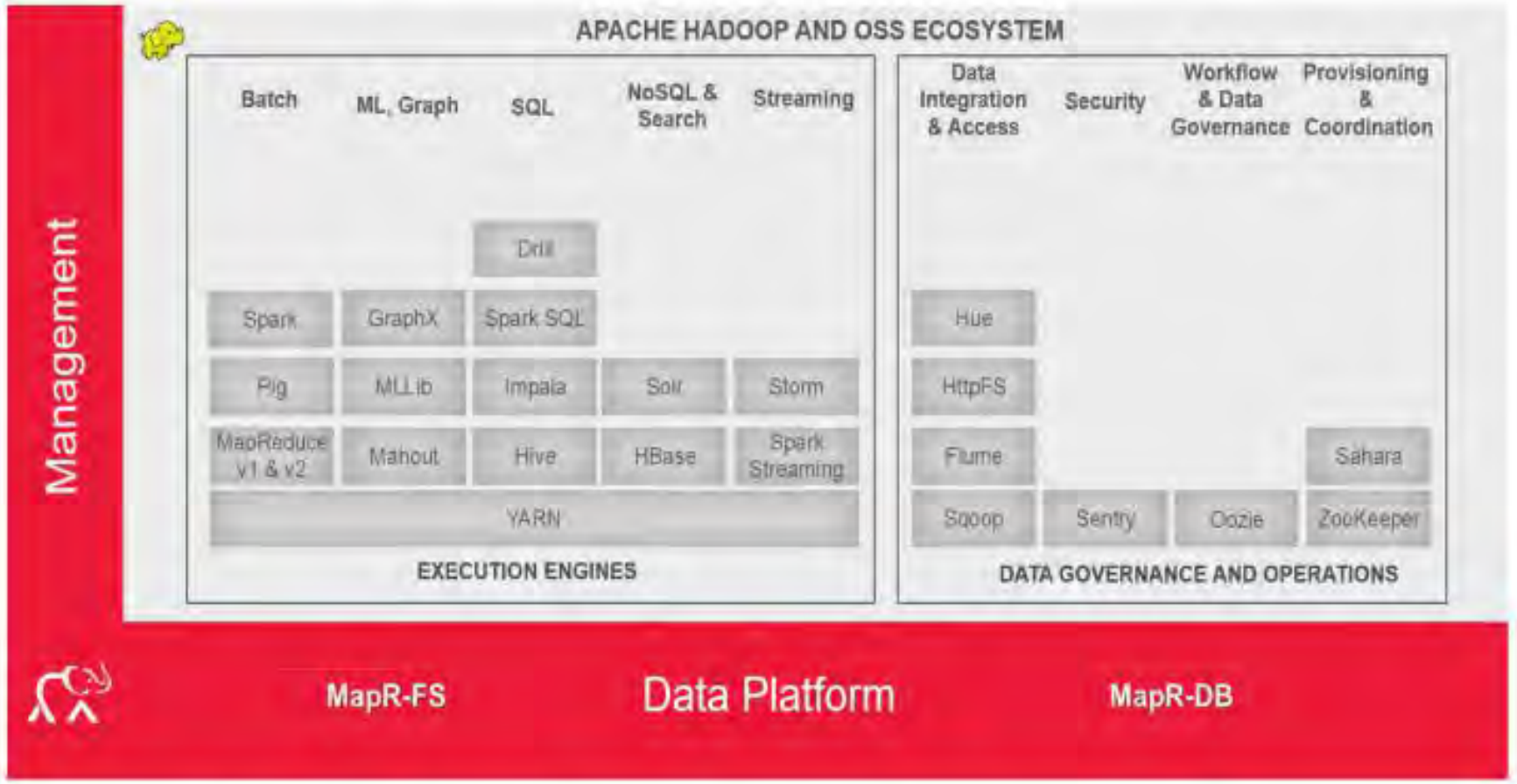


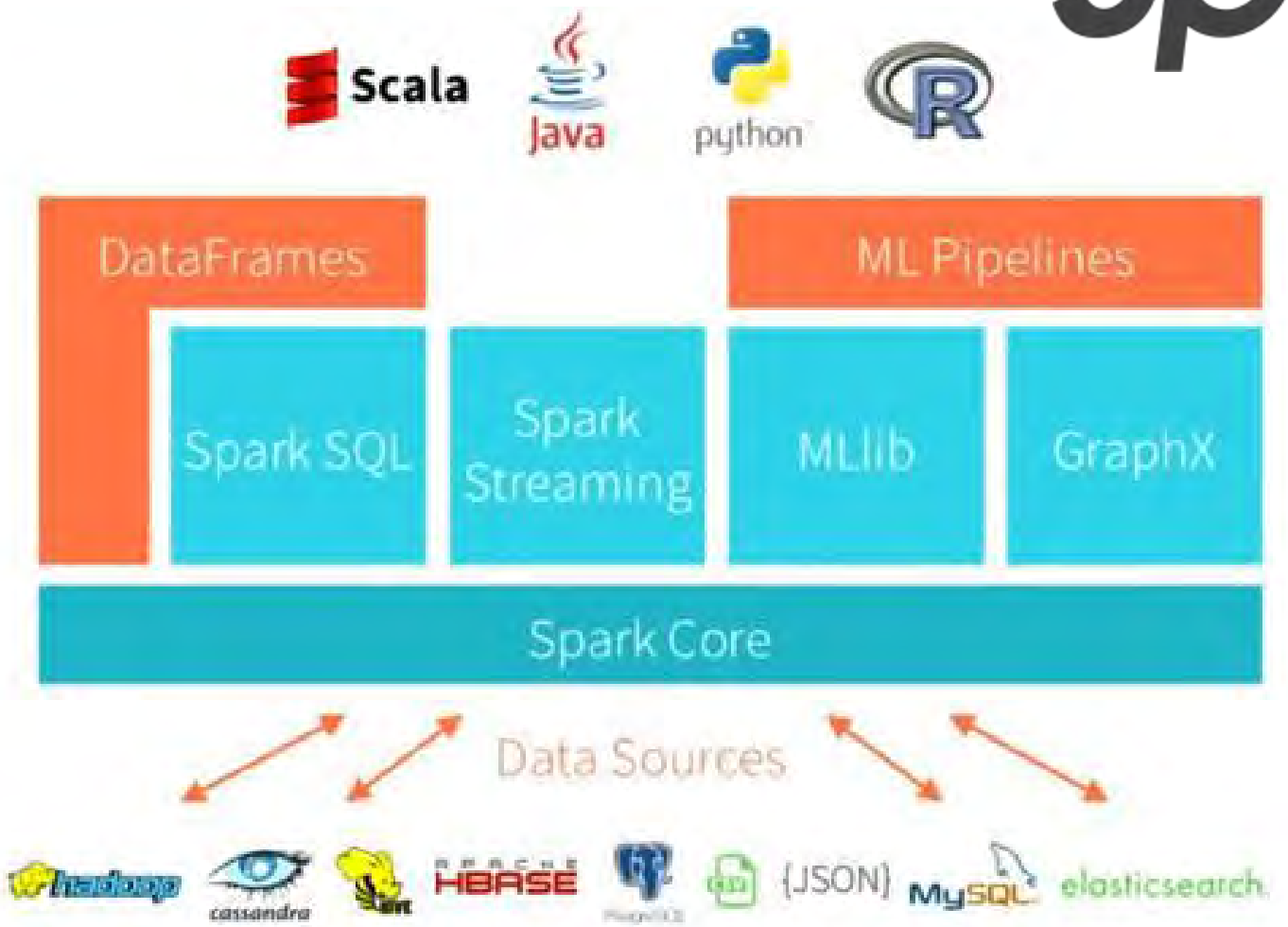
Figure 1. *The MapR Distribution including Apache Hadoop*

Big Data Platforms – Spark

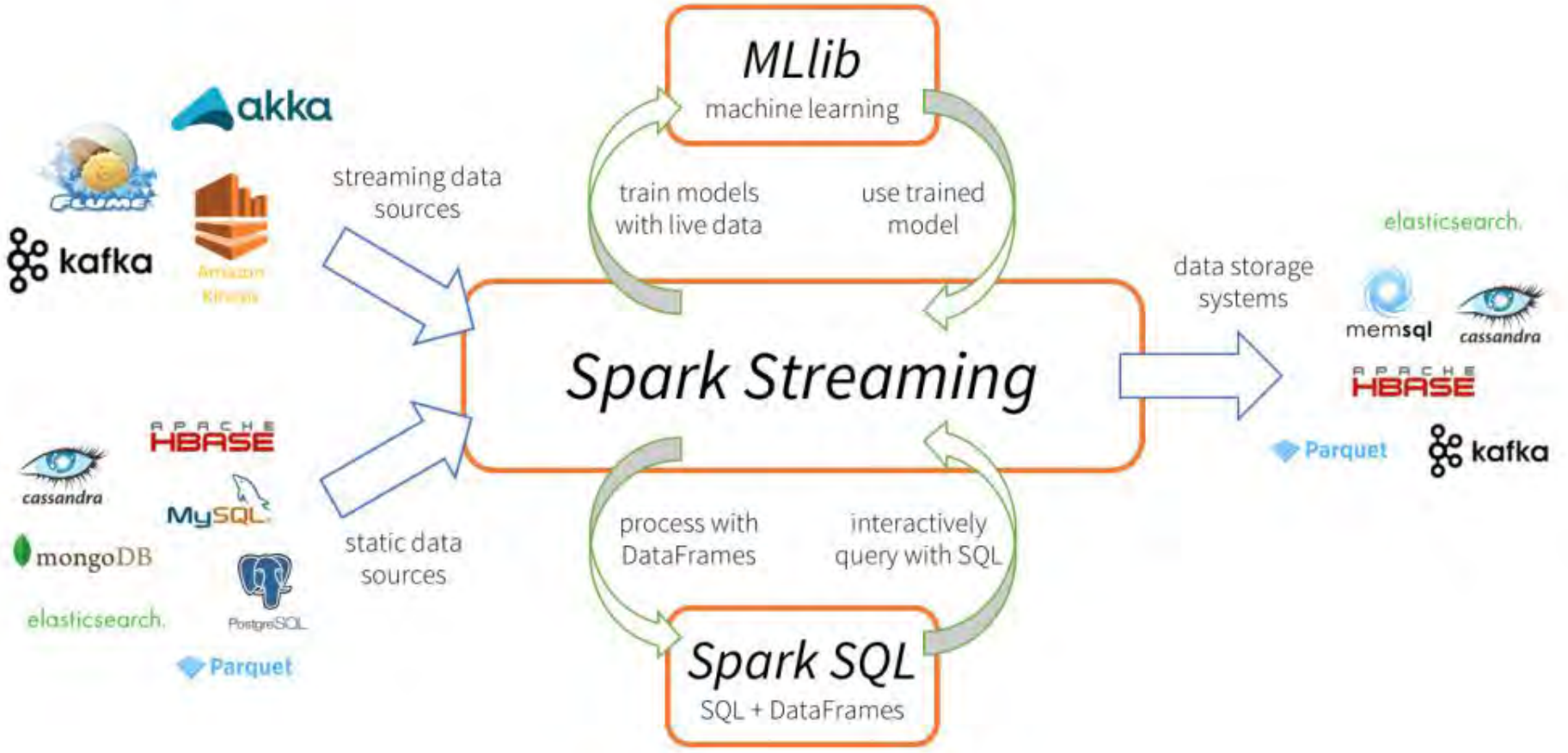
- Open Source processing engine built around speed, ease of use & sophisticated analytics for large-scale data
- In memory…100x faster than MapReduce
- Runs on Hadoop, Standalone, Cloud
- Can access HDFS, Cassandra, S3
- Connectors to Oracle

- SQL, DataFrames, DataSets, Streaming, ML, GraphX
- Largest open source big data project

Spark


Scala
Java
python
DataFrames
ML Pipelines
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core
Data Sources
HBASE
{JSON}
MySQL
elasticsearch

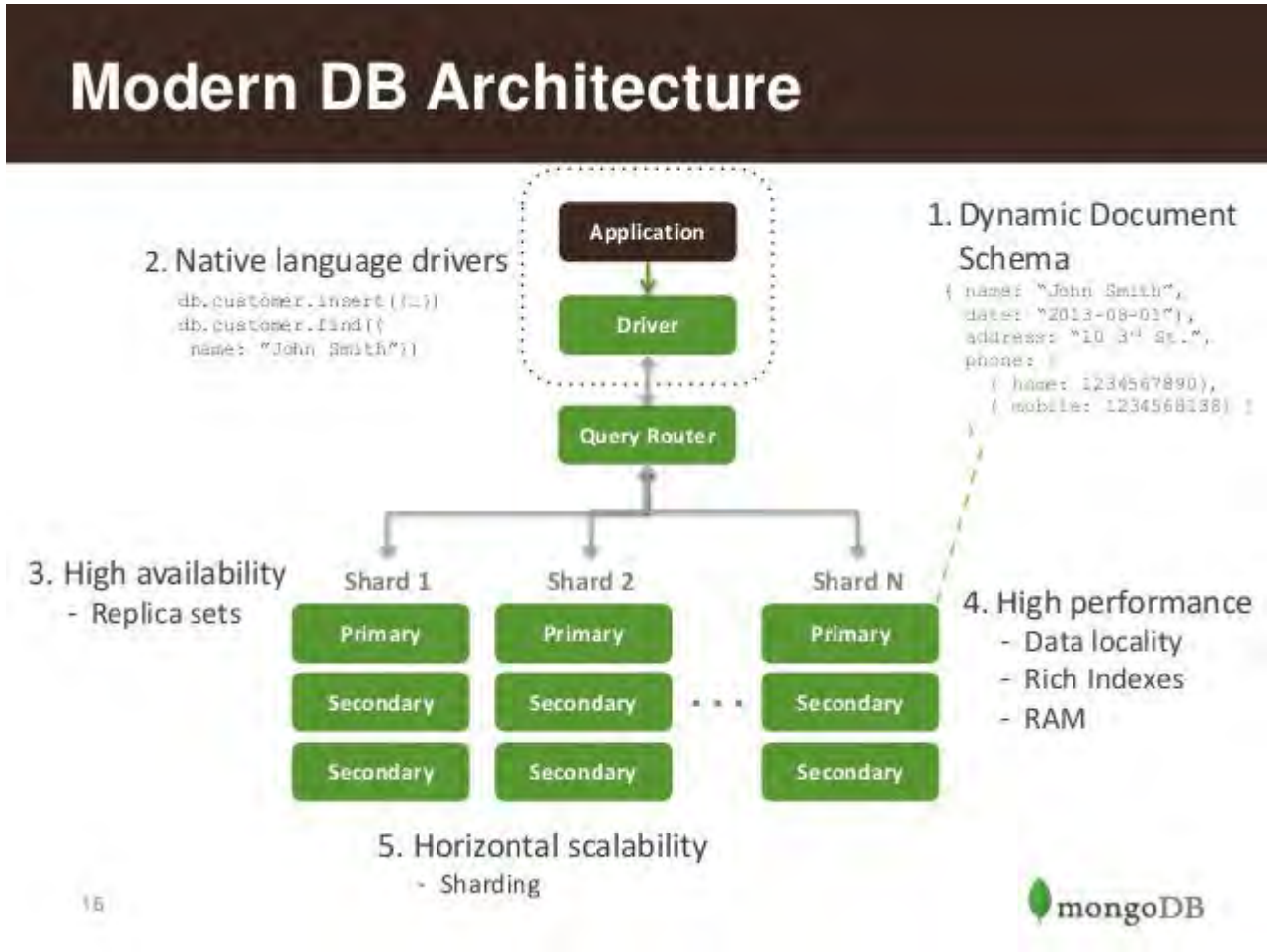
Spark


MLlib
machine learning
train models with live data
use trained model
Spark Streaming
Spark SQL
SQL + DataFrames
process with DataFrames
interactively query with SQL
streaming data sources
static data sources
data storage systems
akka
kafka
Amazon Kinesis
cassandra
HBASE
MySQL
mongoDB
elasticsearch
PostgreSQL
Parquet
memsql
cassandra
HBASE
Parquet
kafka
elasticsearch

Big Data Platforms – NoSQL Mongo

- Storing, retrieving & managing document-oriented information
- Unstructured
- JSON, XML, other
- Subclass of key-value store
- Aligned with modern programming languages
- Mongo:
  - Data model flexibility
  - Elastic scalability
  - High availability

Big Data Platforms – NoSQL Mongo



Big Data Platforms – NoSQL Cassandra

- Non-relational database
- Fault Tolerant & Highly Available yet decentralized
- Performant
- Massively/Linear Scalable
- Easy data distribution across multiple data centers and cloud availability zones

- Tunable Consistency…

Big Data Platforms – NoSQL Cassandra

What is Apache Cassandra

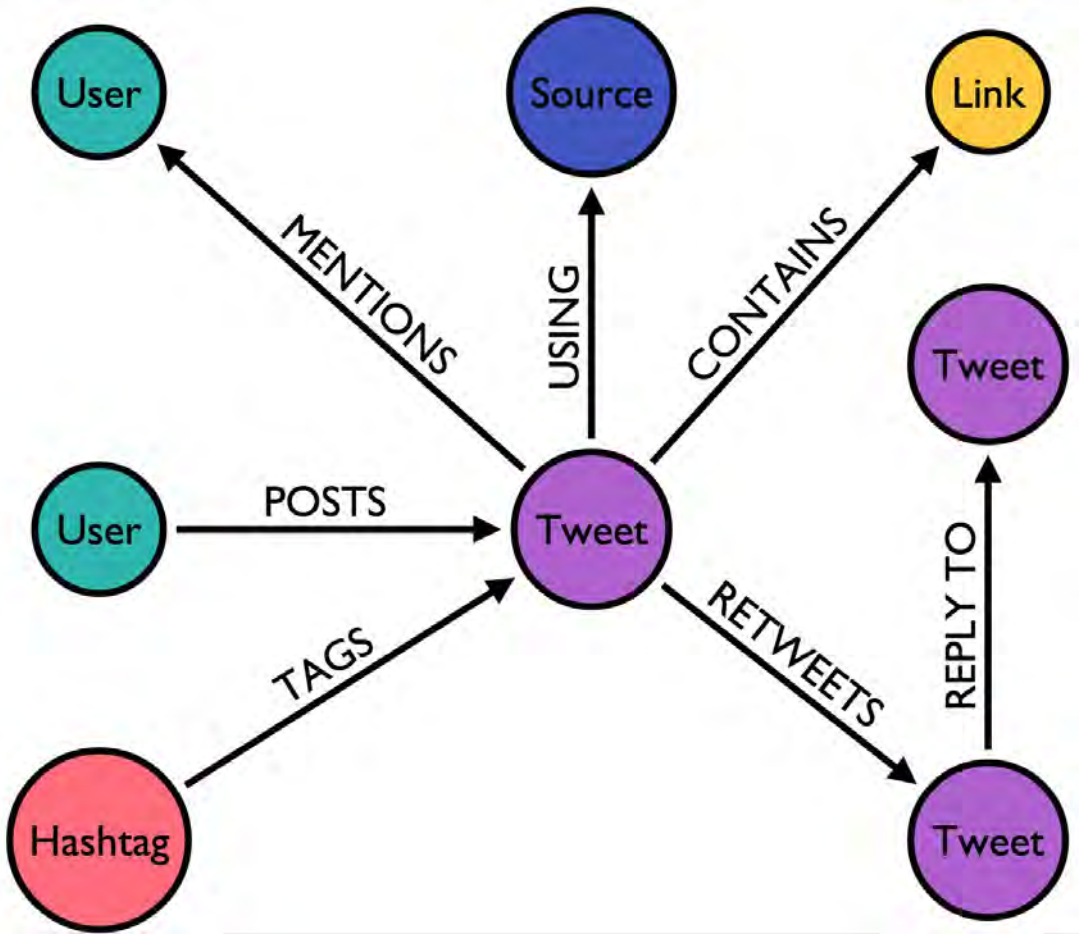
DATASTAX

- Masterless Architecture with read/write anywhere design
- Continuous Availability with no single point of failure
- Multi-Data Center and Zone support
- Flexible data model for unstructured, semi-structured and structured data
- Linear scalable performance with online expansion (scale-out and scale-up)
- Security with integrated authentication
- Operationally simple
- CQL - Cassandra Query Language

Big Data Platforms – Graph neo4j

- Online database system that operates on a graph data model…leveraging data relationships & highly connected data
  - Graph Storage
  - Graph Processing Engine
- Graph is composed of two elements:
  - Node: represents an entity (person, place, thing)
  - Relationship: represents how two nodes are associated
- Neo4j:
  - Native Graph Storage & Processing
  - Highly Performant Read & Write Scalability

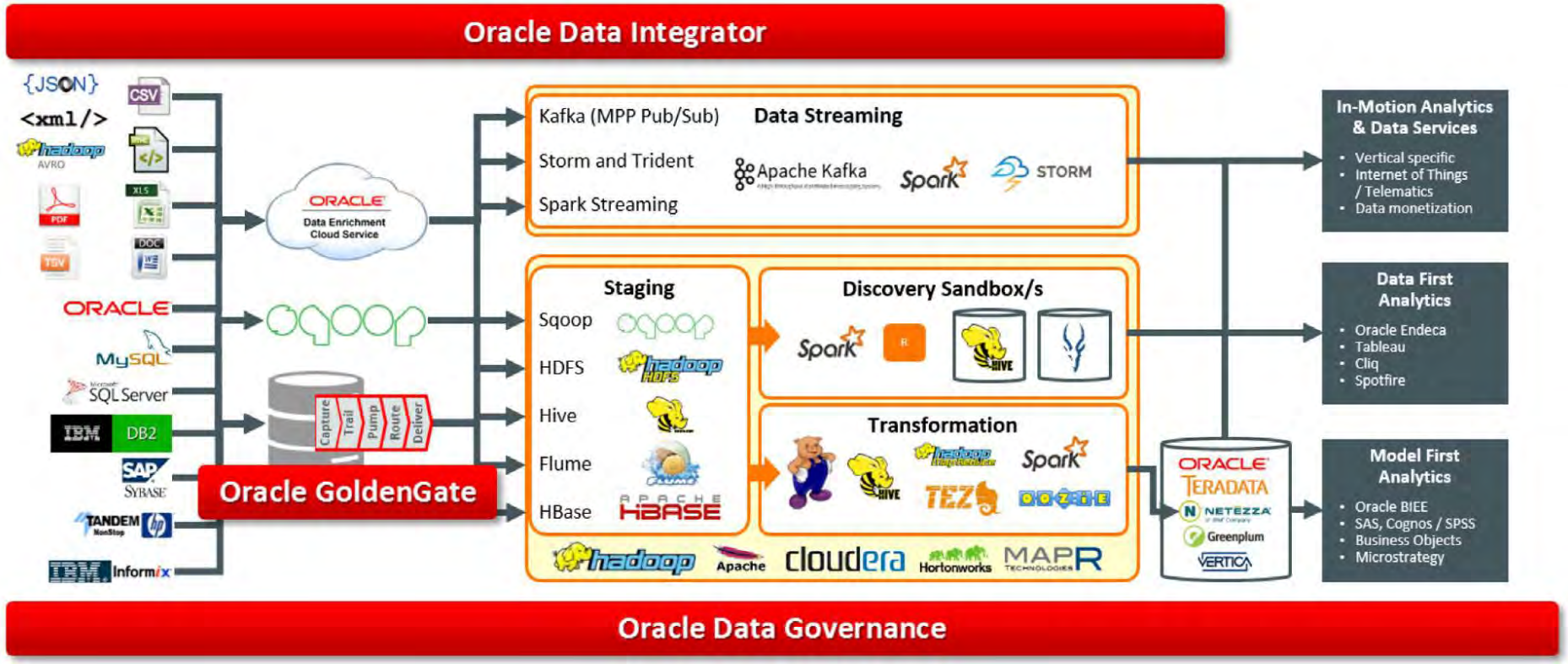
Big Data Platforms – Graph neo4j



Big Data - Integration

- On-Prem
- Cloud
- Hybrid

- Integration
  - Data
  - Application
  - Analytics



Big Data – Integration tools

- Golden Gate/ODI?
- Informatica?

- Talend
- Pentaho
- Gluent – Available?

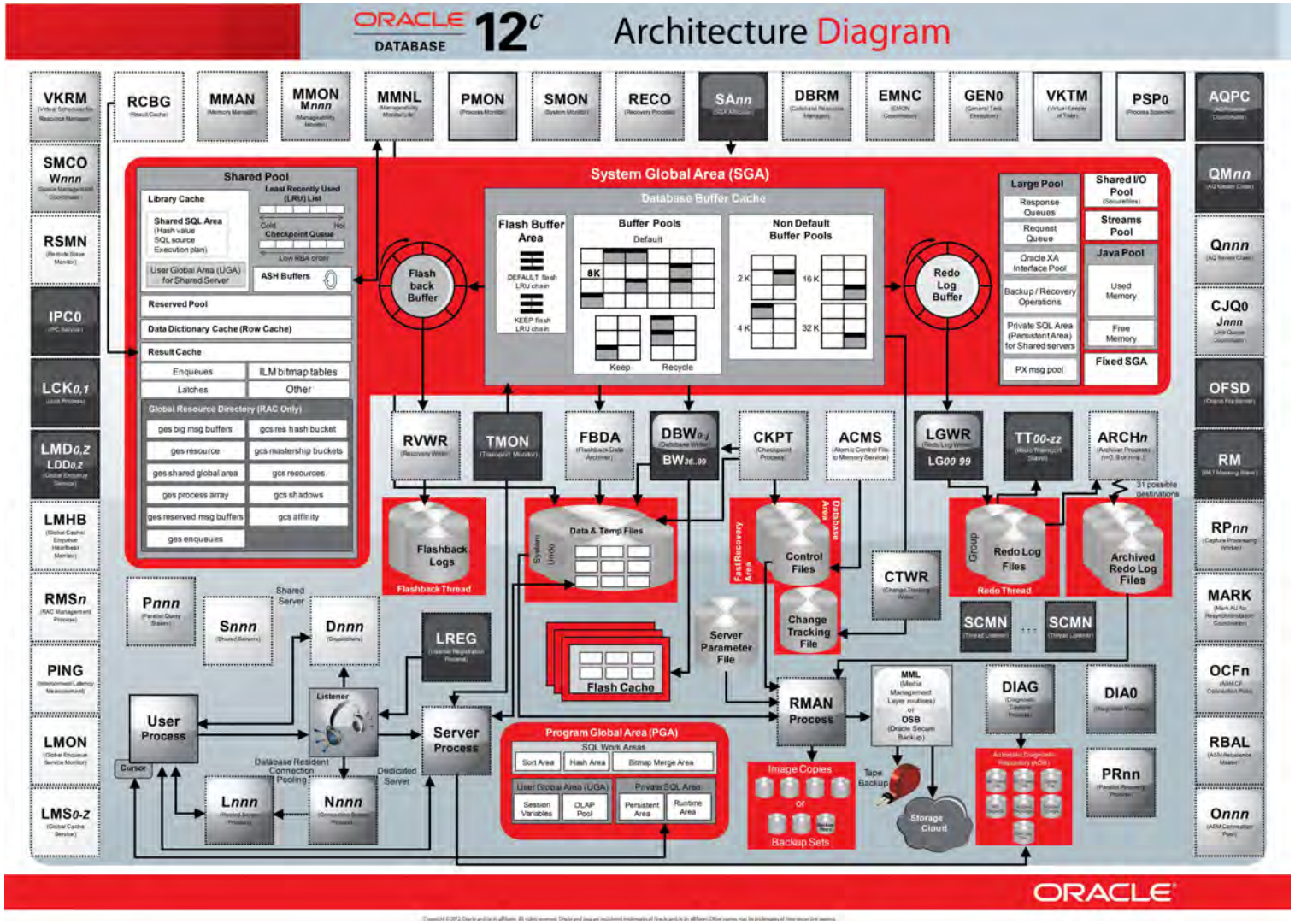
- aws Data Pipeline
- aws Glue…

Oracle DBA

- Confusing?
- Too many moving parts?
- Too much going on very quickly?

- Is there a method to the madness…

Oracle


ORACLE DATABASE 12c
Architecture Diagram
VKRM
RCBG
MMAN
MMON
MMNL
PMON
SMON
RECO
SAnn
DBRM
EMNC
GEN0
VKTM
PSP0
AQPC
SMCO
Wnnn
RSMN
IPC0
LCK0,1
LMD0,Z
LDD0,Z
LMHB
RMSn
PING
LMON
LMS0-Z
QMnn
Qnnn
CJQ0
Jnnn
OFSD
RM
RPnn
MARK
OCFn
RBAL
Onnn
System Global Area (SGA)
Shared Pool
Library Cache
Shared SQL Area
(Hash value
SQL source
Execution plan)
User Global Area (UGA)
for Shared Server
Least Recently Used
(LRU) List
Checkpoint Queue
ASH Buffers
Reserved Pool
Data Dictionary Cache (Row Cache)
Result Cache
Enqueues
ILM bitmap tables
Latches
Other
Global Resource Directory (RAC Only)
ges big msg buffers
gcs res hash bucket
ges resource
gcs mastership buckets
ges shared global area
gcs resources
ges process array
gcs shadows
ges reserved msg buffers
gcs affinity
ges enqueues
Database Buffer Cache
Flash Buffer
Area
DEFAULT flash
LRU chain
KEEP flash
LRU chain
Buffer Pools
Default
8K
Keep
Recycle
Non Default
Buffer Pools
2K
16K
4K
32K
Flash
back
Buffer
Redo
Log
Buffer
Large Pool
Response
Queues
Request
Queue
Oracle XA
Interface Pool
Backup / Recovery
Operations
Private SQL Area
(Persistent Area)
for Shared servers
PX msg pool
Shared I/O
Pool
Streams
Pool
Java Pool
Used
Memory
Free
Memory
Fixed SGA
RVWR
TMON
FBDA
DBW0-j
BW36-99
CKPT
ACMS
LGWR
LG00-99
TT00-zz
ARCHn
31 possible
destinations
Flashback
Logs
Flashback Thread
Data & Temp Files
System
Undo
Fast Recovery
Area
Control
Files
Database
Area
CTWR
Change
Tracking
File
Group
Redo Log
Files
Redo Thread
Archived
Redo Log
Files
SCMN
SCMN
Pnnn
Snnn
Shared
Server
Dnnn
LREG
Flash Cache
Server
Parameter
File
RMAN
Process
MML
(Media
Management
Layer routines)
or
OSB
(Oracle Secure
Backup)
DIAG
DIA0
Listener
User
Process
Cursor
Server
Process
Database Resident
Connection
Pooling
Dedicated
Server
Lnnn
Nnnn
Program Global Area (PGA)
SQL Work Areas
Sort Area
Hash Area
Bitmap Merge Area
User Global Area (UGA)
Session
Variables
OLAP
Pool
Private SQL Area
Persistent
Area
Runtime
Area
Image Copies
or
Backup Sets
Tape
Backup
Storage
Cloud
PRnn
ORACLE

Oracle DBA…Cont'd

- Rapidly evolving toolset
- Support of various components still an issue
- Continue to learn new technologies

- Oracle DBA is in a unique role!
- Be ready to succeed in a polyglot environment where various storage & analytics platforms will be used!

Best of Breed - Repeat

- Polyglot Persistence
  - Pick the right storage & analytics engine for the right use case
  - Hybrid

- The purpose of data platforms remains the same:
  - Store data
  - Retrieve data
  - Analyze/Process data
  - How efficiently we do this depends on the platform!

Top Big Data Use Cases

- Enterprise Data Hub/Lake
- Data Warehouse Offload
- ETL/ELT Offload
- Stream processing
- AI/ML

Real-World Use Case 1

- Oil & Gas
  - Oracle RAC:
    - Streaming/IOT data from field was being fed into Oracle
    - Due to volume & velocity of data, had severe performance issues
    - Data was rarely retrieved/used due to heavy cost
  - Apache Hadoop on MapR:
    - On-Prem MapR based Hadoop Cluster setup
    - Was able to ingest TB's of Streaming/IOT data from field in Real-Time
    - Real-Time stream processing, analytics allowed for anomaly detection
    - Bi-directional replication from/to Hadoop & Oracle
    - Tremendous Cost savings!
    - Apache Hadoop on MapR

    - https://www.wsj.com/articles/fracking-2-0-shale-drillers-pioneer-new-ways-to-profit-in-era-of-cheap-oil-1490894501?emailToken=JRrydv16Y3iXhNMzacwyzlQjbagOBKrTAwuSN3DDPkWJuGbUpeas3b5wn9qwp26iXwN86s9B8GcuXnjNhy9yRsifmqI6kFHhdmNU65bKlAaLN03D2UmLea9F6viNrng1s/EC

Real-World Use Case 2

- Healthcare Analytics
  - Oracle RAC:
    - Processed 100’s of millions of records of supply chain data from hospitals on a nightly/batch basis to generate & provide advanced analytics to customers
    - Process took ~12hrs and application was unavailable when process was underway
    - License cost was high & many features weren’t used due to cost
    - Could not scale due to hardware & other restrictions
  - Apache Spark/DataBricks on AWS/S3:
    - Data Warehouse Workload offloaded to Apache Spark on AWS/S3
    - Real-Time Machine Learning now possible
    - Many new features are now in use!
    - Could scale to any number!
    - Tremendous time & Cost savings!

Summary

- Big Data Overview
- Big Data Platforms
- Promise of Hadoop & **Spark**
- Hybrid
  - **Polyglot**: Platforms
  - Hybrid: Cloud vs On-Prem

References

- https://www.oracle.com
- https://cloud.oracle.com/bigdata

- http://spark.apache.org/
- https://www.cloudera.com
- https://www.cloudera.com/products/enterprise-data-hub.html
- https://mapr.com/
- https://hortonworks.com/

- https://www.mongodb.com/
- https://neo4j.com/developer/graph-database/
- http://cassandra.apache.org/

Q&A
Please complete evaluations!

Tom S. Reddy

dataREDDY LLC
tom.reddy@datareddy.com

Oracle OpenWorld 2017 #OOW17

SessionID: SUN5698

The age of Big Data

Big Data for Oracle Database Professionals

Tom S. Reddy

tom.reddy@datareddy.com

dataREDDY

Oracle OpenWorld 2017 #OOW17

SessionID: SUN5698