aioug
ALL INDIA ORACLE USERS GROUP
THE POWER TO CONNECT & CONQUER

Transition from an Oracle DBA to Big Data architect

Saurabh K. Gupta
@saurabhkg

Who am I?

- Database Leader, Data and Analytics at GE
- 10 years of experience in data engineering, architecture, Oracle technologies
- Authored couple of books with Packt Publishing
  - Oracle Advanced PL/SQL Developer Professional Guide
  - Advanced Oracle PL/SQL Developer's Guide – Second Edition (12c)
- Twitter @saurabhkg

Why I'm here?

- As a classical Oracle Database administrators, you know how to deal with fat data sets already. Big data is little different as more than its size, what matters is the variety and velocity.
- Businesses are staking a lot to find the data nuggets out of noisy heaps. There is a lot that DBAs can contribute in this shift. Not just the data availability, but DBAs can transform themselves into data architects by stepping out of classical database administration skills.
- This session will focus on skill areas that can help Oracle DBAs to emerge as Big Data DBAs. The talk will cover the overview of big data ecosystem, key Big Data technologies and what DBAs can leverage from their current skill set to focus on big data DBA.

Agenda

- Big Data – making sense out of nonsense
- How to design a Big Data solution?
- Big Data solution spectrum
- Build you Big Data team

Big Data – Making Sense out of Nonsense

Big Data – Making Sense out of Nonsense

- Structured and unstructured data that augments a business on daily basis –
  - Large volumes of data
  - At a *relative* velocity
  - With *relative* variety
  - Can reveal nuggets of information
- Information is more important than Data
  - What we do with the data
- New term; not the concept
  - Data gathering, storage, and analysis has been for a while

Big Data – what Industry thinks?

- *"Data really powers everything that we do."* – Jeff Weiner, LinkedIn.
- *"You can have data without information, but you cannot have information without data."* - Daniel Keys Moran
- *“Data beats emotions.”* – Sean Rad, founder of Ad.ly
- *“Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world.”* – Atul Butte, Stanford
- *“Torture the data, and it will confess to anything.”* – Ronald Coase, Economics, Nobel Prize Laureate

How much we love Data

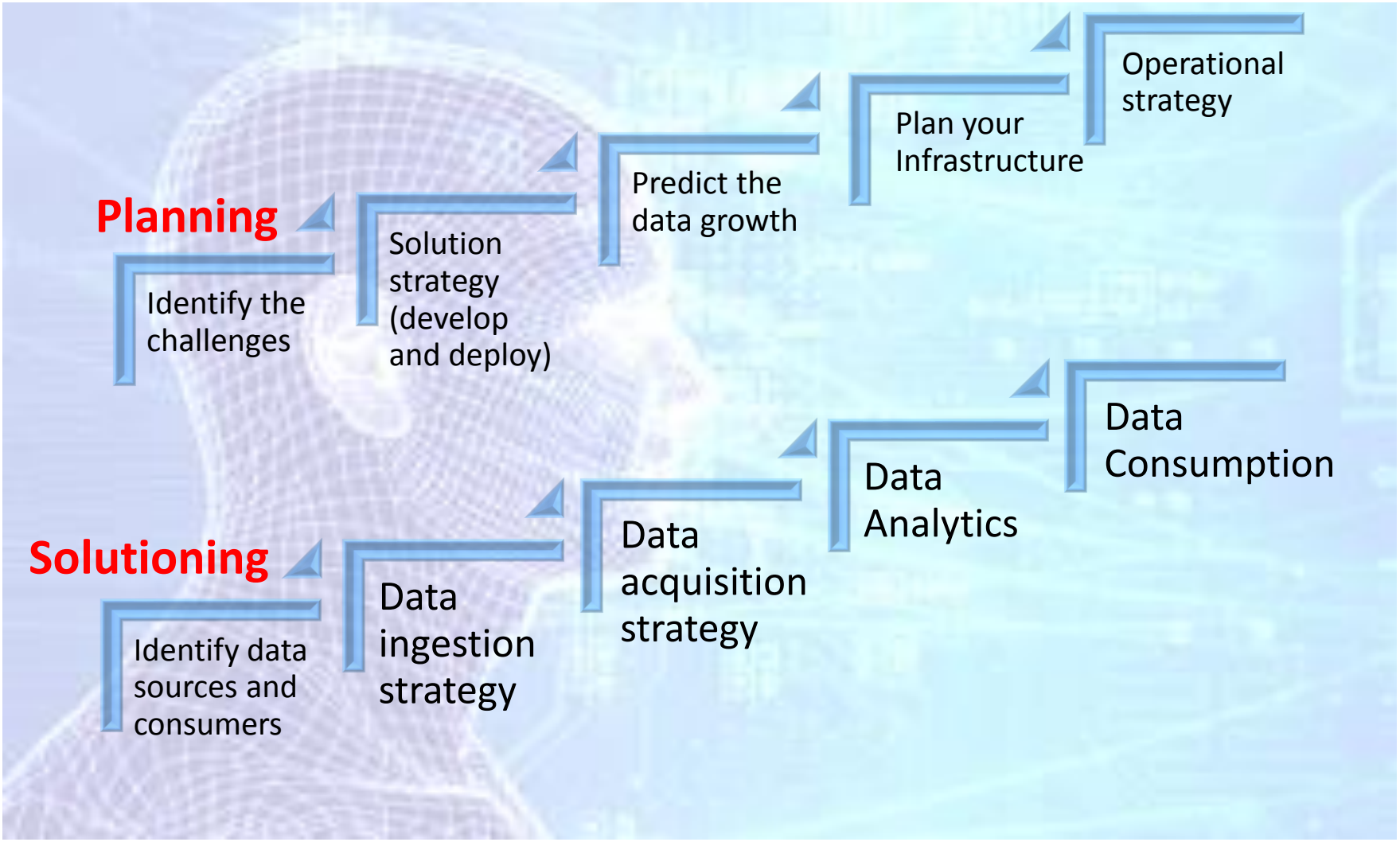
- Facebook (http://newsroom.fb.com/company-info/)
  - 1.18 billion daily active users on average for Sep'16
- Twitter (https://about.twitter.com/company)
  - 313M active monthly users; 82% active users on mobile
- Instagram (https://www.instagram.com/press/)
  - 4.2 Billion likes daily; 30+ Billion in ~5 years
- Google searches
  - 57,115 Google searches in 1 second
- Digital universe will grow to 44 zettabytes (Trillion GB)

How much we love Data

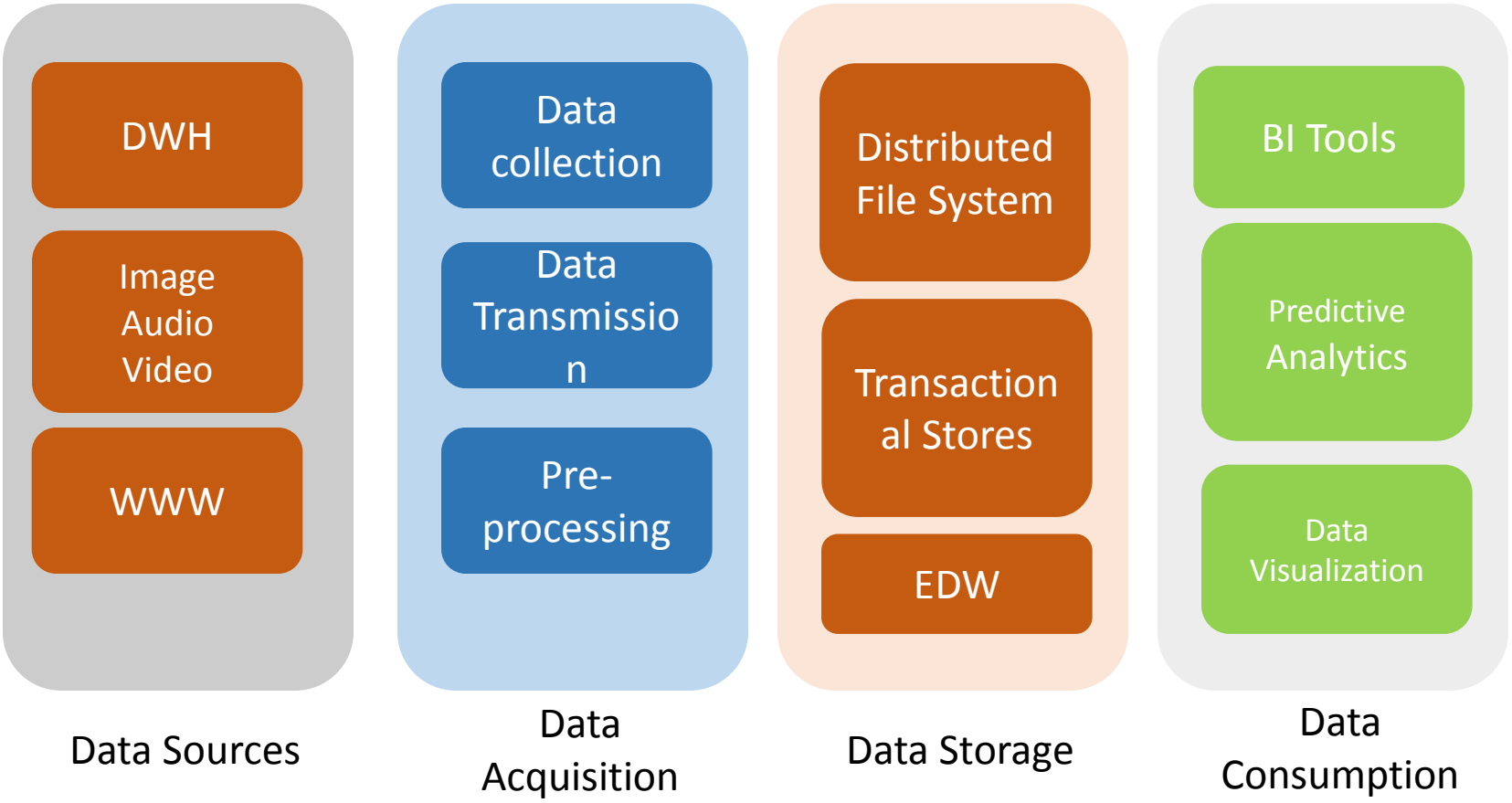
- *"Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone."*
- *- IBM*

How to design a Big Data solution

Plan and develop Big Data solution



Big Data solution spectrum



https://hadoopecosystemtable.github.io/

Data Acquisition and Ingestion

- Design strategy for data collection, data transmission, data pre-processing
- Understand the nature of data
  - Data gen rate, volume, batch or stream
- Data ingestion tools - Sqoop, Flume, Kafka, Storm
- Web crawling tools (Apache Nutch and open-source)
- Oracle GoldenGate for Big Data 12c (https://www.oracle.com/goldengate/big-data/index.html)

Data Acquisition and Ingestion

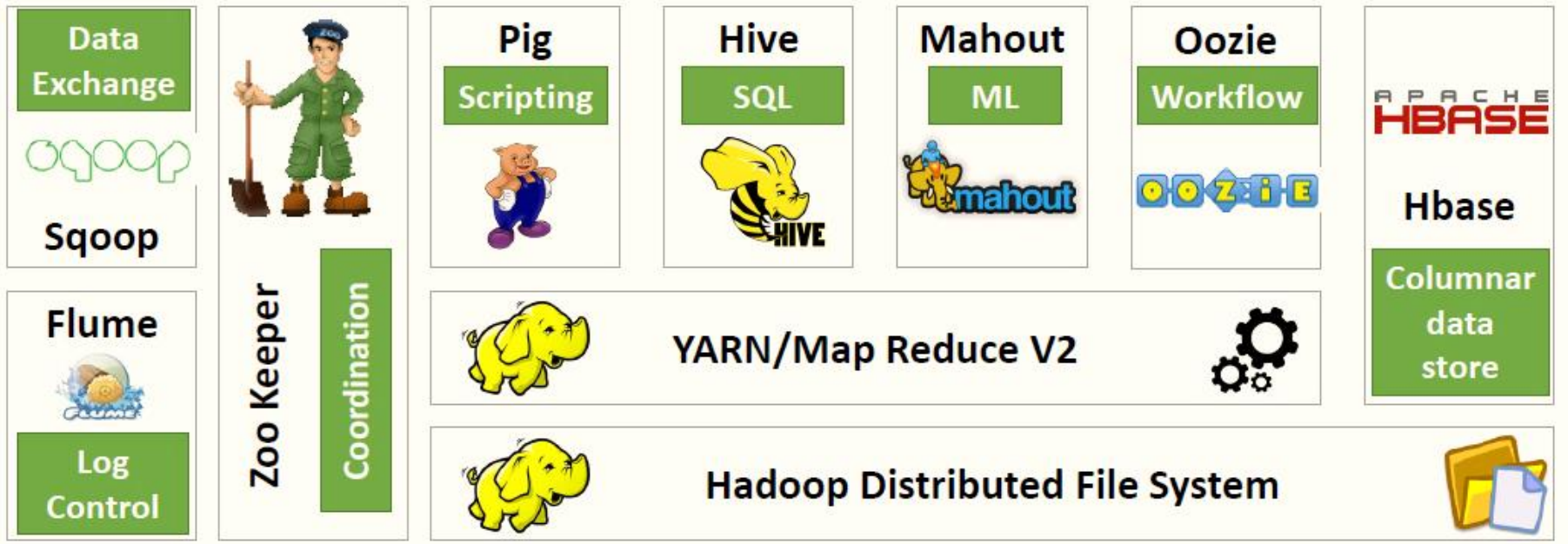
- Flume - Distributed system for collecting and aggregating log data, and writing it to HDFS. Simple, flexible, and highly available. Tightly integrated with Hadoop.
- Sqoop - Provides two way replication between Apache Hadoop and RDBMS. Supports snapshots and incremental updates.
- Kafka - distributed publish-subscribe messaging system. Hadoop is a consumer of Kafka.
- Storm - distributed computation based event-processing system. Often referred as real-time Hadoop. Storm cluster coordinates with Zookeeper.
- Others – Chukwa, Scribe, Samza,

Data Storage

Apache Hadoop –

- Framework used for multiple-node processing of data
  - Provides both distributed storage and distributed processing of very large data sets
- Scalable platform for processing large batches of data very fast; High degrees of parallelism
  - Master slave architecture

Apache Hadoop



*...evolving**

Evolution of the Hadoop Ecosystem

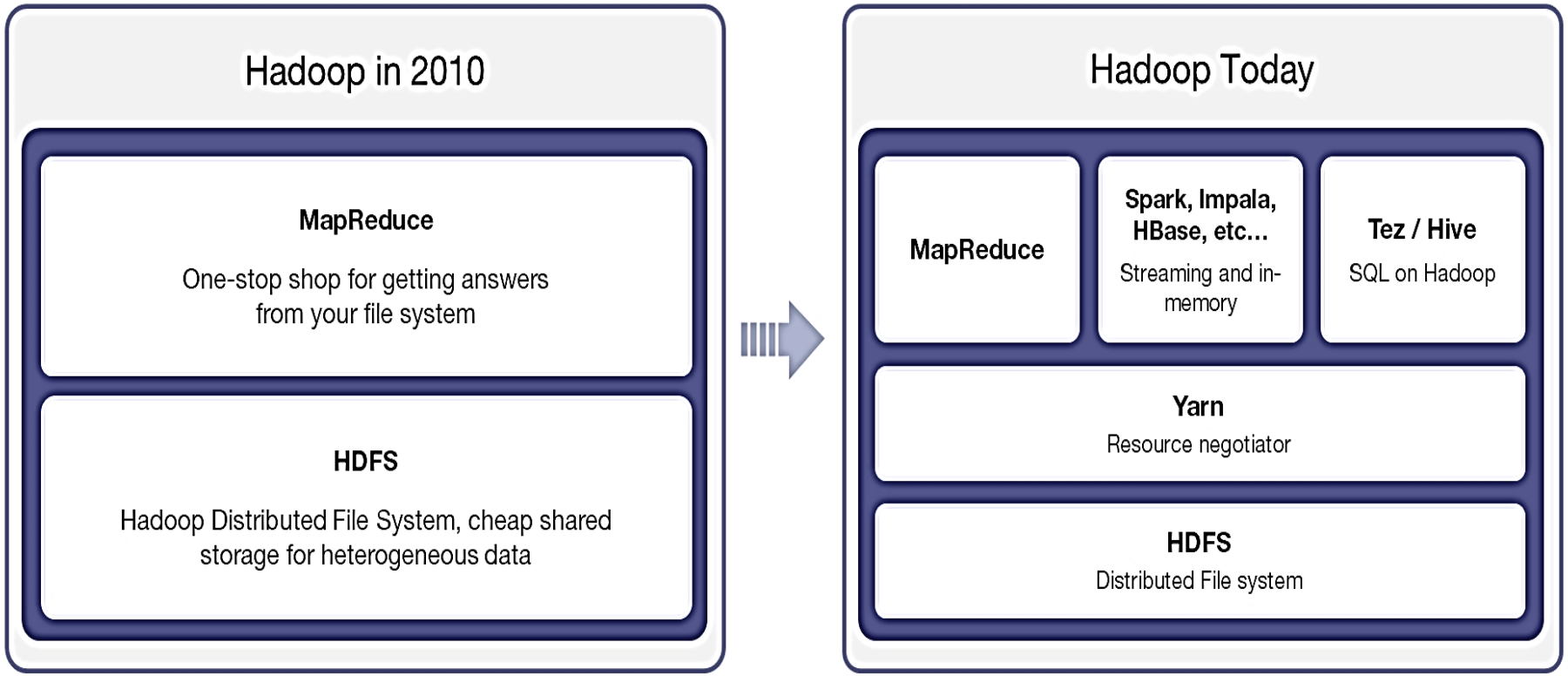


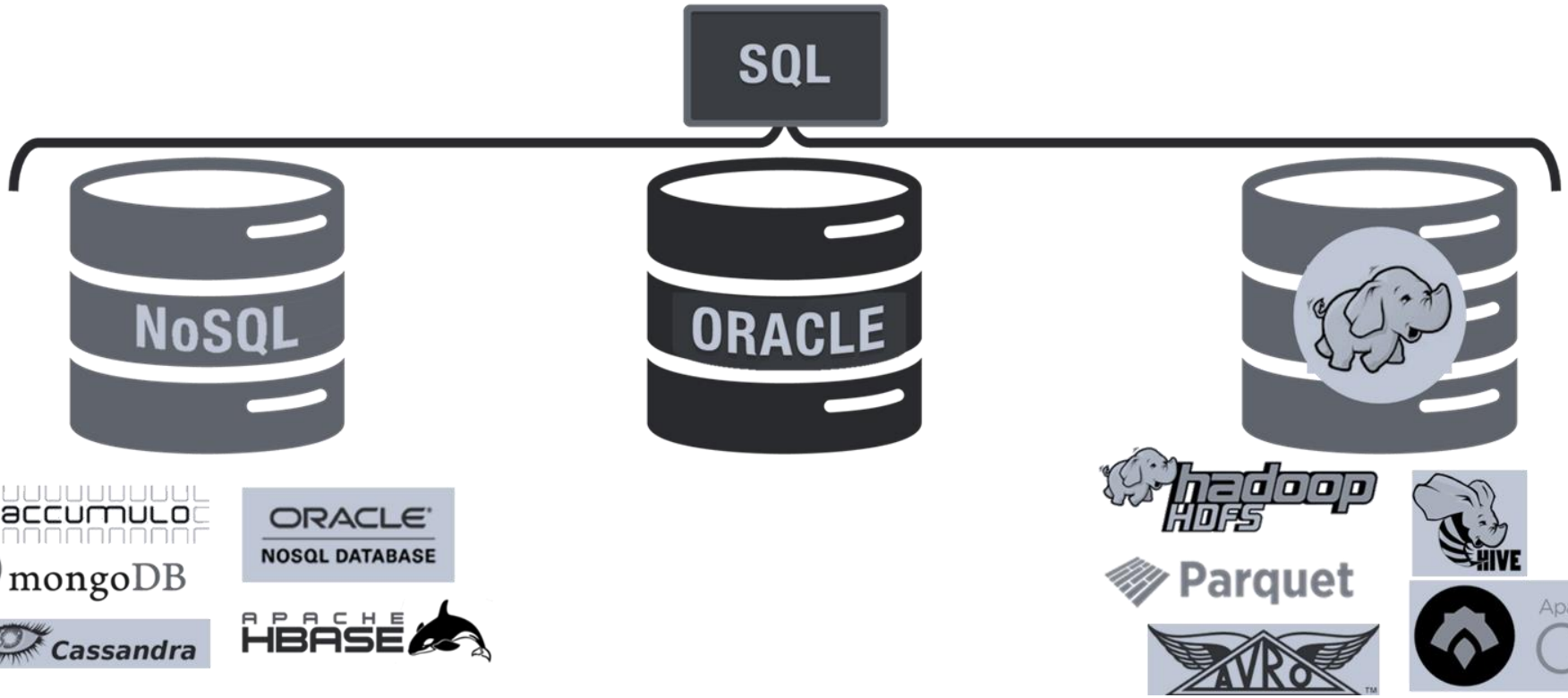
Image Source - http://www.marklogic.com/blog/tdwi-hadoop-readiness-assessment-and-guide/

Data scrubbing with Pig and Hive

- Data warehousing capability on top of Hadoop
- HiveQL provides familiarity SQL folks
- Uses MapReduce for execution
- Enable data mining on large volumes of data

- Dataflow scripting language
- Uses MapReduce for execution
- Pig interpreter submits the jobs to Hadoop cluster

Oracle Big Data SQL



Oracle Big Data SQL

- Powerful, high-performance SQL on Hadoop
  - Full Oracle SQL capabilities on Hadoop
  - SQL query processing local to Hadoop nodes
- Simple data integration of Hadoop and Oracle Database
  - Single SQL point-of-entry to access all data
  - Scalable joins between Hadoop and RDBMS data
- Optimized hardware
  - High-speed Infiniband network between Hadoop and Exadata

Transactional data-stores

- RDBMS do not scale with massive volumes of data
- NoSQL main characteristics is it's non-adherence to relational database concepts of CODD
- Focus on scalability, performance, high availability
  - ACID properties are not always guaranteed
  - No joins, less complex, no constraints

Transactional data-stores

- Availability of data is more important than data consistency (BASE)
- Relations are addressed at application level
- Go by CAP theorem
  - Consistency, Availability, Partition tolerance

Transactional data-stores

- Key-value
  - Oracle NoSQL, DynamoDB, Voldermort, Apache Accumulo
- Document-based
  - MongoDB, CouchDB
- Column-based
  - Apache Cassandra, Apache Hbase
- Graph-based
  - Neo4J, InfoGrid
- Relational
  - Apache Kudu

Data Analytics

- Convergence layer where volume, velocity, and variety transform into Value
- Structured data analysis – Data mining, Inferential statistics
- Text Analysis – Natural Language Processing, Text mining, Opinion mining, categorization
- Web Analytics – Web mining, Web usage analysis
- Network Analytics – Social media based
- Mobile Analytics – location based mining
- Multi media analytics – event detection and prediction

Build your Big Data team

Administrator	ETL Developer	Data Architect	Big Data Architect	Data Scientist
• Hadoop admins • Information security • DevOps	• Implement ETL/ELT flow • Sqoop, Flume, ETL tools, Stream processing	• Data modeling • Hadoop • ETL design • Data analytics	• Develop core applications using NoSQL, Spark, MapReduce • Data processing • Data Visualization strategy	• Statistical techniques • Machine learning • R, Python, Perl

Questions?


рахмат
danke
謝謝
ngiyabonga
spas
teşekkür ederim
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
tapadh leat
спасибо
faafetai lava
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
thank you
gracias
хвала
kiitos
dankie
nanni
nandri
dhanyavad
mauruuru
aćiū
asante
manana
bedankt
bayarlalaa
gràçie
hvala
köszönöm
dankon
akun
obrigada
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
chokrane
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
obrigado
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
najis tuke
tanemirt
grazie
terima kasih
rahmet
kam sah hamnida
rahmat
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci