Big Data Analytics

Amit Juggurnath

Analytics and Blockchain Solutions Lead
SADC

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. | Confidential – Oracle Internal/Restricted/Highly Restricted

Meet Taylor—
32 years old; lives in San Francisco

Coffee Drinker

Recently purchased an expensive espresso machine

Favorite brand is Durham Denim. Visits their website often for new products.

Sporty; likes hiking, camping, and outdoor activities

It’s a sunny Sunday in San Francisco and nearly 75 degrees.

Current Practices

Predefined

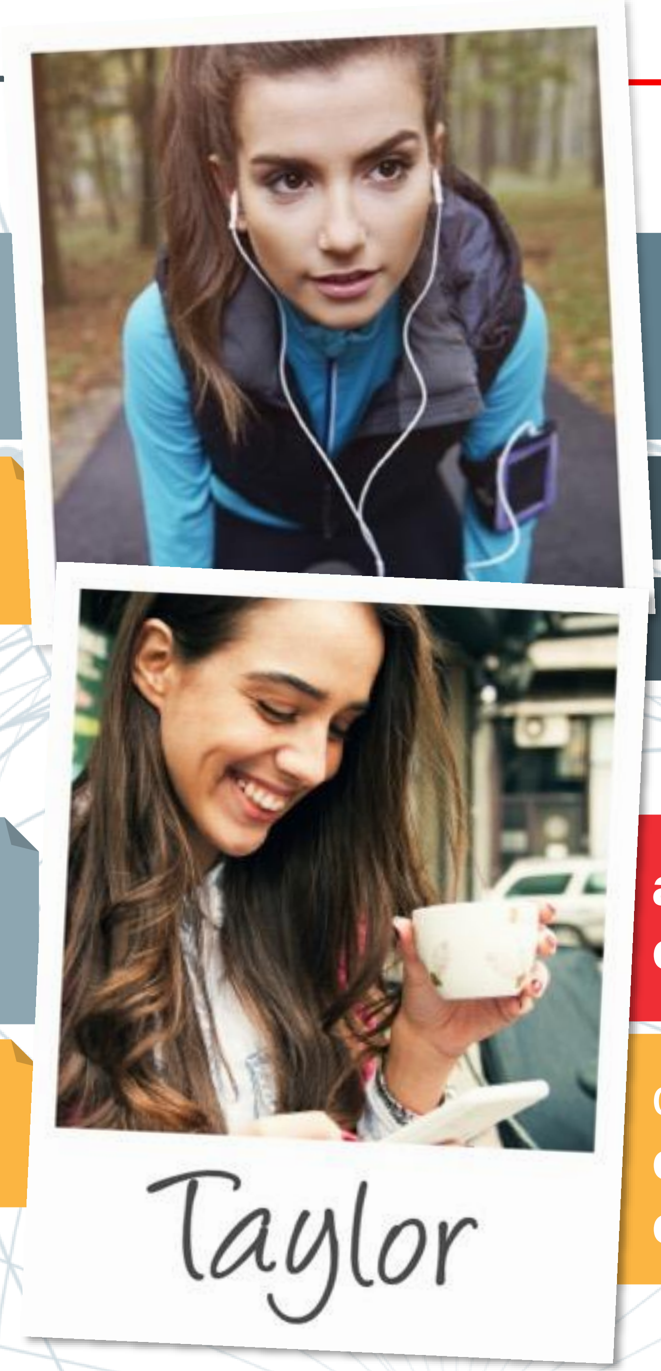
Performs preset scenarios such as “buy 6, get the 7th coffee free” and…

on Sunday, Taylor receives an offer for a **free coffee**.

Predictive

Predicts Taylor wants coffee based on past transactions and...

on Sunday, Taylor receives an offer for **free coffee**.



Adaptive Intelligence

Processes insight from Taylor’s digital footprint, recent transactions, and…

social activity… weather… location…

others like her at this moment…

and real-time actions. Then…

anticipates she needs a cooler drink, **considers** her high price threshold, and…

on Sunday, Taylor is offered a **free iced coffee,** and an **up-sell offer** for **premium coffee beans**.

ORACLE

Oracle Confidential. Internal Use Only. Not to be distributed.

Big Data Analysis Techniques

1. Association rule learning
2. Classification tree analysis
3. Genetic algorithms
4. Machine learning
5. Regression analysis
6. Sentiment analysis
7. Social network analysis

https://www.firmex.com/thedealroom/7-big-data-techniques-that-create-business-value/

1. Association rule learning

Are people who purchase tea more or less likely to purchase carbonated drinks?

- Association rule learning is a method for discovering interesting correlations between variables in large databases. It was first used by major supermarket chains to discover interesting relations between products, using data from supermarket point-of-sale (POS) systems.
- Association rule learning is being used to help:
  - place products in better proximity to each other in order to increase sales
  - extract information about visitors to websites from web server logs
  - analyze biological data to uncover new relationships
  - monitor system logs to detect intruders and malicious activity
  - identify if people who buy milk and butter are more likely to buy diapers

2. Classification tree analysis

Which categories does this document belong to?

- Statistical classification is a method of identifying categories that a new observation belongs to. It requires a training set of correctly identified observations – historical data in other words.
- Statistical classification is being used to:
  - automatically assign documents to categories
  - categorize organisms into groupings
  - develop profiles of students who take online courses

3. Genetic algorithms

Which TV programs should we broadcast, and in what time slot, to maximize our ratings?

- Genetic algorithms are inspired by the way evolution works – that is, through mechanisms such as inheritance, mutation and natural selection. These mechanisms are used to "evolve" useful solutions to problems that require optimization.
- Genetic algorithms are being used to:
  • schedule doctors for hospital emergency rooms
  • return combinations of the optimal materials and engineering practices required to develop fuel-efficient cars
  • generate "artificially creative" content such as puns and jokes

4. Machine learning

Which movies from our catalogue would this customer most likely want to watch next, based on their viewing history?

- Machine learning includes software that can learn from data. It gives computers the ability to learn without being explicitly programmed, and is focused on making predictions based on known properties learned from sets of "training data."
- Machine learning is being used to help:
  - distinguish between spam and non-spam email messages
  - learn user preferences and make recommendations based on this information
  - determine the best content for engaging prospective customers
  - determine the probability of winning a case, and setting legal billing rates

5. Regression analysis

How does your age affect the kind of car you buy?

- At a basic level, regression analysis involves manipulating some independent variable (i.e. background music) to see how it influences a dependent variable (i.e. time spent in store). It describes how the value of a dependent variable changes when the independent variable is varied. It works best with continuous quantitative data like weight, speed or age.
- Regression analysis is being used to determine how:
  • levels of customer satisfaction affect customer loyalty
  • the number of supports calls received may be influenced by the weather forecast given the previous day
  • neighbourhood and size affect the listing price of houses
  • to find the love of your life via online dating sites

6. Sentiment analysis

How well is our new return policy being received?

- Sentiment analysis helps researchers determine the sentiments of speakers or writers with respect to a topic.
- Sentiment analysis is being used to help:
  - improve service at a hotel chain by analyzing guest comments
  - customize incentives and services to address what customers are really asking for
  - determine what consumers really think based on opinions from social media

7. Social network analysis

How many degrees of separation are you from Tom Cruise?

- Social network analysis is a technique that was first used in the telecommunications industry, and then quickly adopted by sociologists to study interpersonal relationships. It is now being applied to analyze the relationships between people in many fields and commercial activities. Nodes represent individuals within a network, while ties represent the relationships between the individuals.
- Social network analysis is being used to:
  - see how people from different populations form ties with outsiders
  - find the importance or influence of a particular individual within a group
  - find the minimum number of direct ties required to connect two individuals
  - understand the social structure of a customer base

Lifecycle of a Data Analytics Project

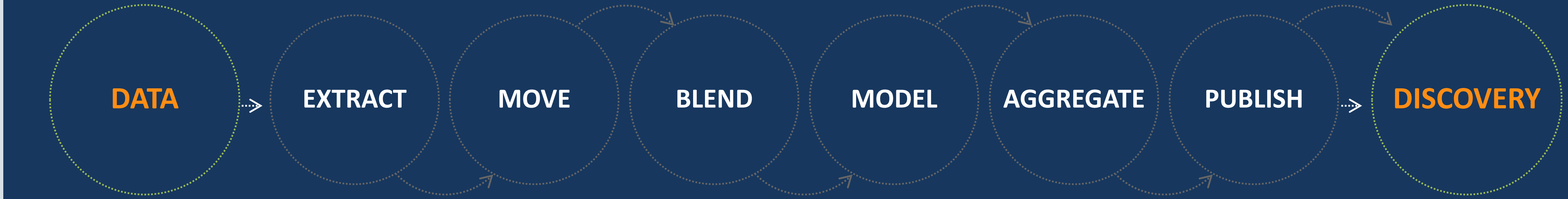
Copyright © 2017, Oracle and/or its affiliates. All rights reserved. | Confidential – Oracle Internal/Restricted/Highly Restricted

Generic Lifecycle of a Data Analytics Project

- Identifying the problem
- Designing the data requirement
- Pre processing data
- Performing analytics over data
- Visualizing data

https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781782163282/5/ch05lvl1sec36/understanding-the-data-analytics-project-life-cycle

HOW THINGS HAVE ALWAYS BEEN DONE



COST AND COMPLEXITY

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

HOW THINGS SHOULD BE DONE.

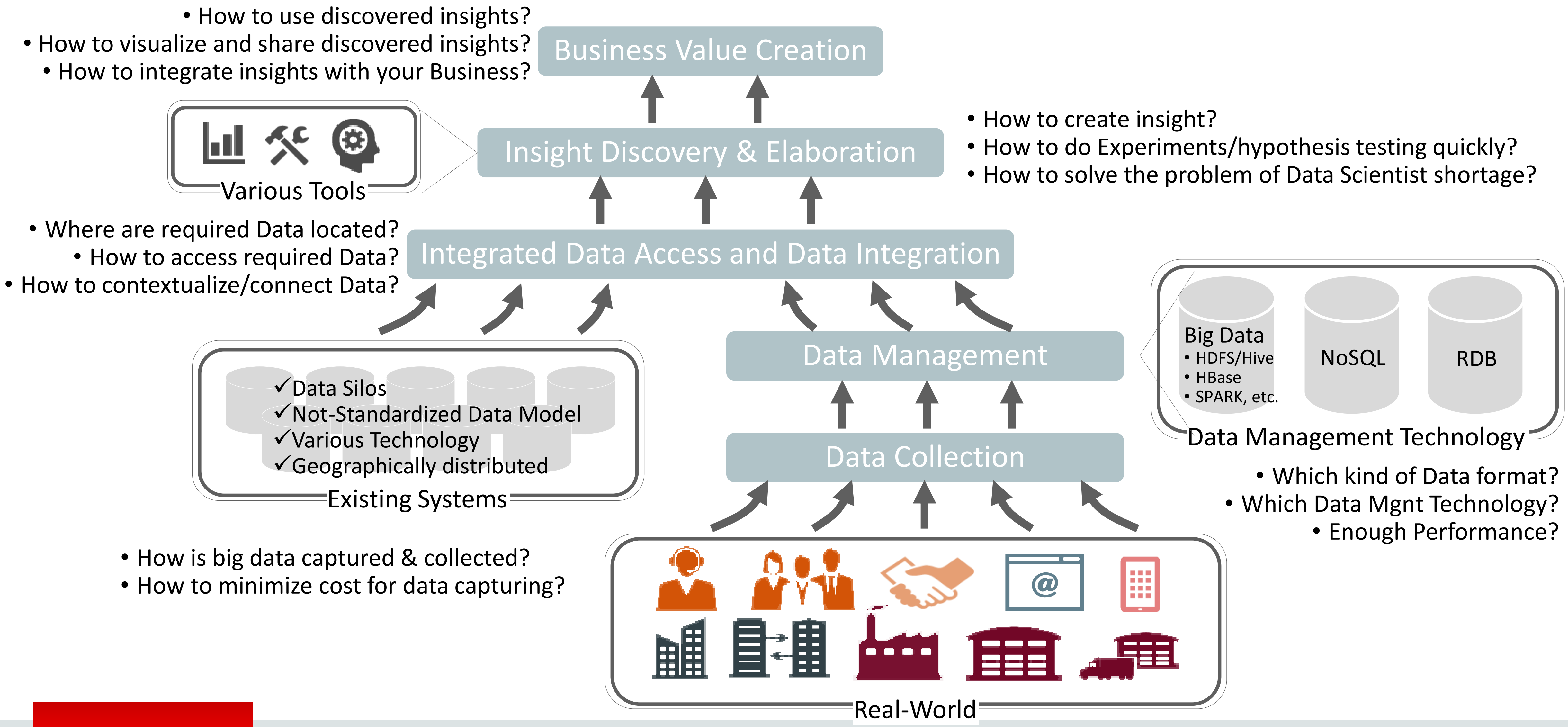
DATA

EMBEDDED MACHINE LEARNING

DISCOVERY

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Challenges around Machine Learning & Big Data



- How to use discovered insights?
- How to visualize and share discovered insights?
- How to integrate insights with your Business?

Business Value Creation

Insight Discovery & Elaboration

Various Tools

- How to create insight?
- How to do Experiments/hypothesis testing quickly?
- How to solve the problem of Data Scientist shortage?

- Where are required Data located?
- How to access required Data?
- How to contextualize/connect Data?

Integrated Data Access and Data Integration

Existing Systems

✓Data Silos
✓Not-Standardized Data Model
✓Various Technology
✓Geographically distributed

Data Management

Data Management Technology

Big Data
- HDFS/Hive
- HBase
- SPARK, etc.

NoSQL

RDB

- Which kind of Data format?
- Which Data Mgnt Technology?
- Enough Performance?

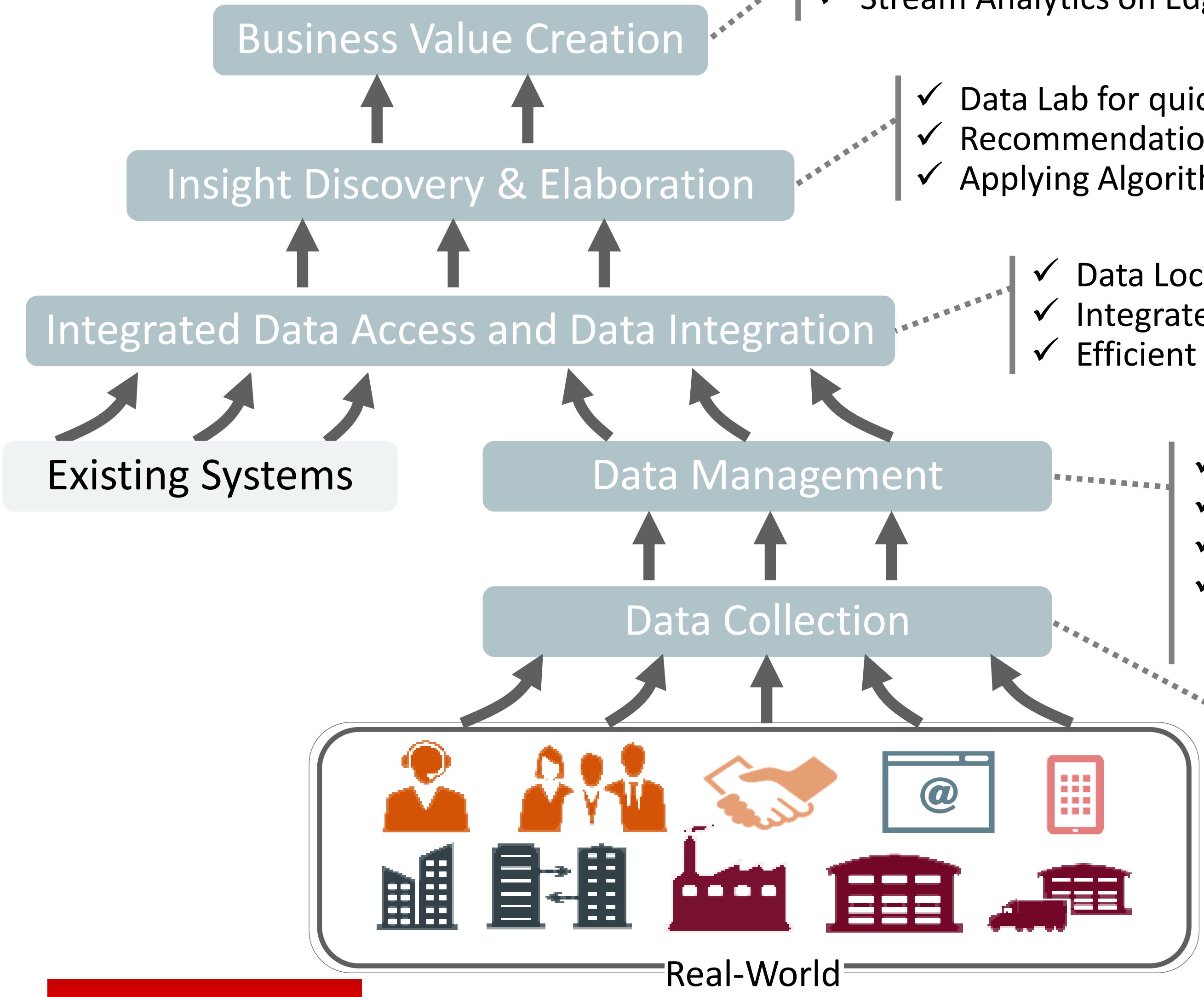
Data Collection

Real-World

- How is big data captured & collected?
- How to minimize cost for data capturing?

Copyright © 2015, Oracle and/or its affiliates. All rights reserved. |

Technology innovations supporting Big Data & Machine Learning



Business Value Creation

Insight Discovery & Elaboration

Integrated Data Access and Data Integration

Existing Systems

Data Management

Data Collection

Real-World

- ✓ Machine Learning Model in Business Apps and BI, to judge/evaluate incoming data
- ✓ Stream Analytics on Edge Computing System, for real-time control optimization

- ✓ Data Lab for quick hypothesis testing execution, with enough sample data
- ✓ Recommendation of Best-Fit Algorithms
- ✓ Applying Algorithms directly to production data, without critical performance impact

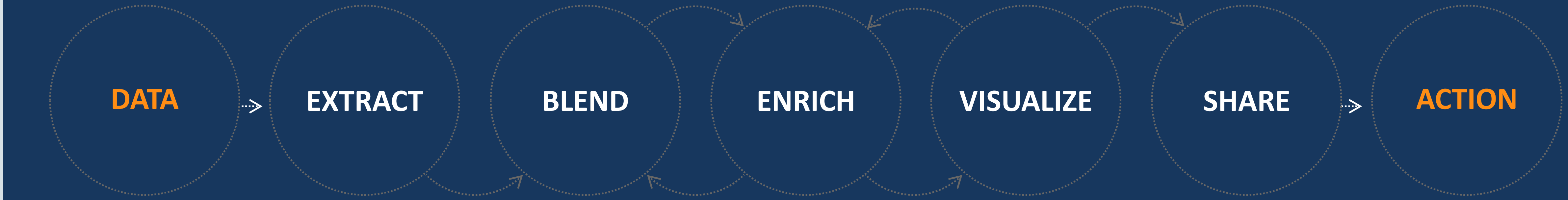
- ✓ Data Location and semantics management (Metadata Mgmt & ontology)
- ✓ Integrated Data Access across systems and technologies
- ✓ Efficient Data Access Method for geo-distributed data

- ✓ Data stored in distributed cheaper storage
- ✓ Performance Improvement with distributed in-memory caches/indexes
- ✓ Column-oriented storage & Columnar Compression
- ✓ Quicker Query Processing by Software-in-Silicon/Software-on-Chip

- ✓ Data Capturing by Cheap & Small Board-Computer
- ✓ Data Capturing by Mobile Devices & Wearable Devices, with Low-Cost App Development Framework & Tools
- ✓ Visualization of Field-Info through Movie & Image Capturing & Analysis
- ✓ Real-Time Embedded Processing Unit

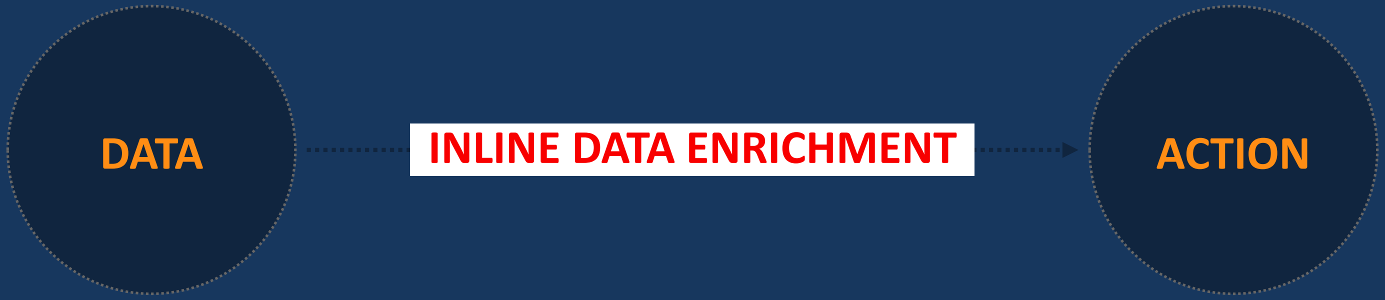
Copyright © 2015, Oracle and/or its affiliates. All rights reserved. |

HOW THINGS HAVE ALWAYS BEEN DONE.



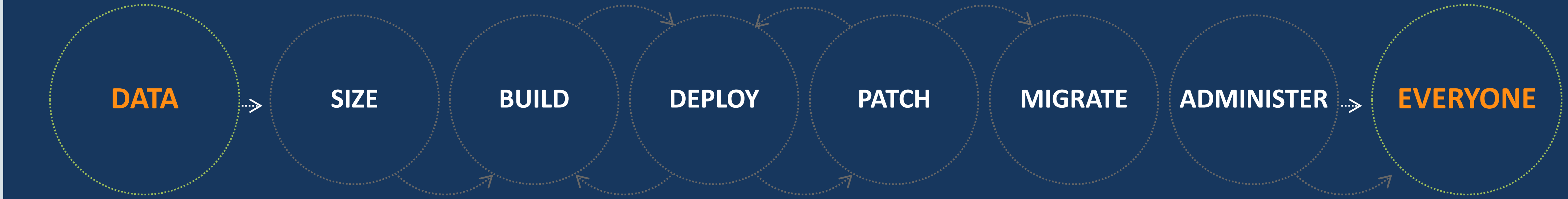
Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

HOW THINGS SHOULD BE DONE.



Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

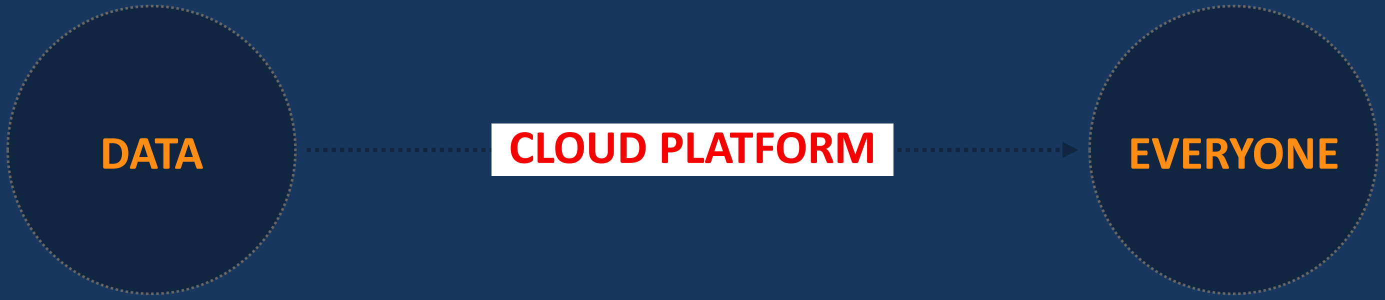
HOW THINGS HAVE ALWAYS BEEN DONE.



REPETITIVE TASKS AND HUMAN ERROR

Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

HOW THINGS SHOULD BE DONE.

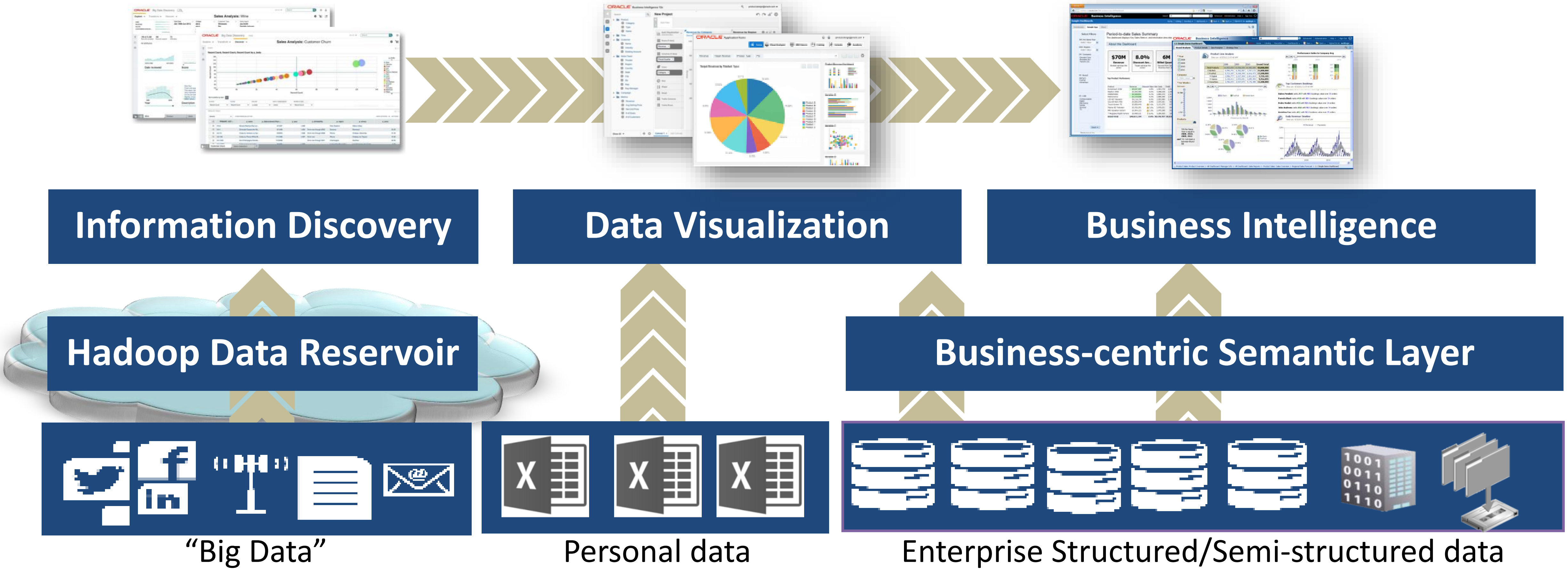


Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Information to Insight

Modern Approach

In Cloud
Hybrid
On Premises



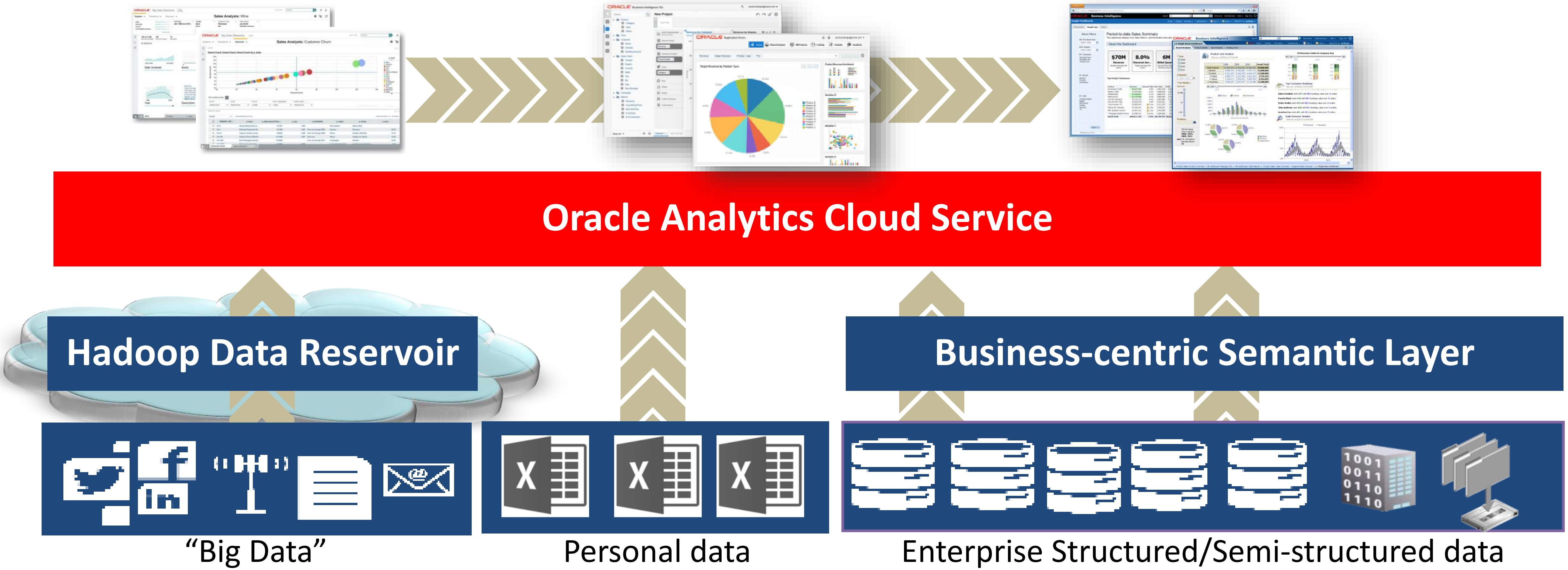
Data Flow through the Enterprise

Modern Approach

Oracle Public Cloud

Hybrid

On Premises



Oracle Analytics Cloud Service

Hadoop Data Reservoir

Business-centric Semantic Layer

“Big Data”

Personal data

Enterprise Structured/Semi-structured data

All Data, Any Size, Any Location

Oracle Analytics Cloud

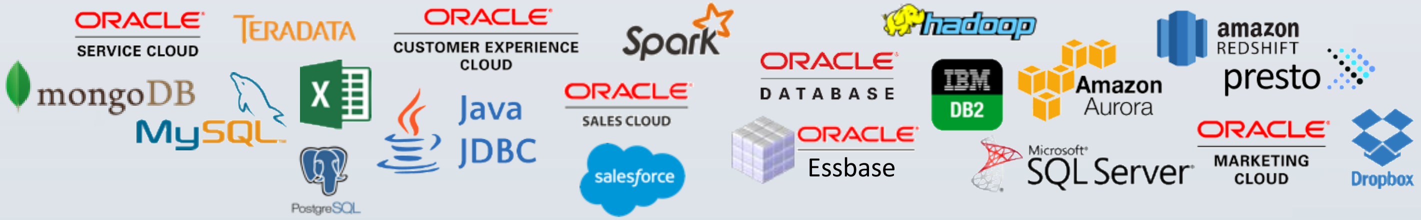
Data Analysis and Collaboration
Explore and discover using natural language, visualization, & storytelling

Data Preparation
Prepare enriched, sharable, & reliable data sets

Data and Model Catalog
One place to collect, search, explore & curate all data, Self Service along side enterprise semantics.

Oracle Autonomous Data Warehouse Cloud

Oracle Big Data Cloud Service



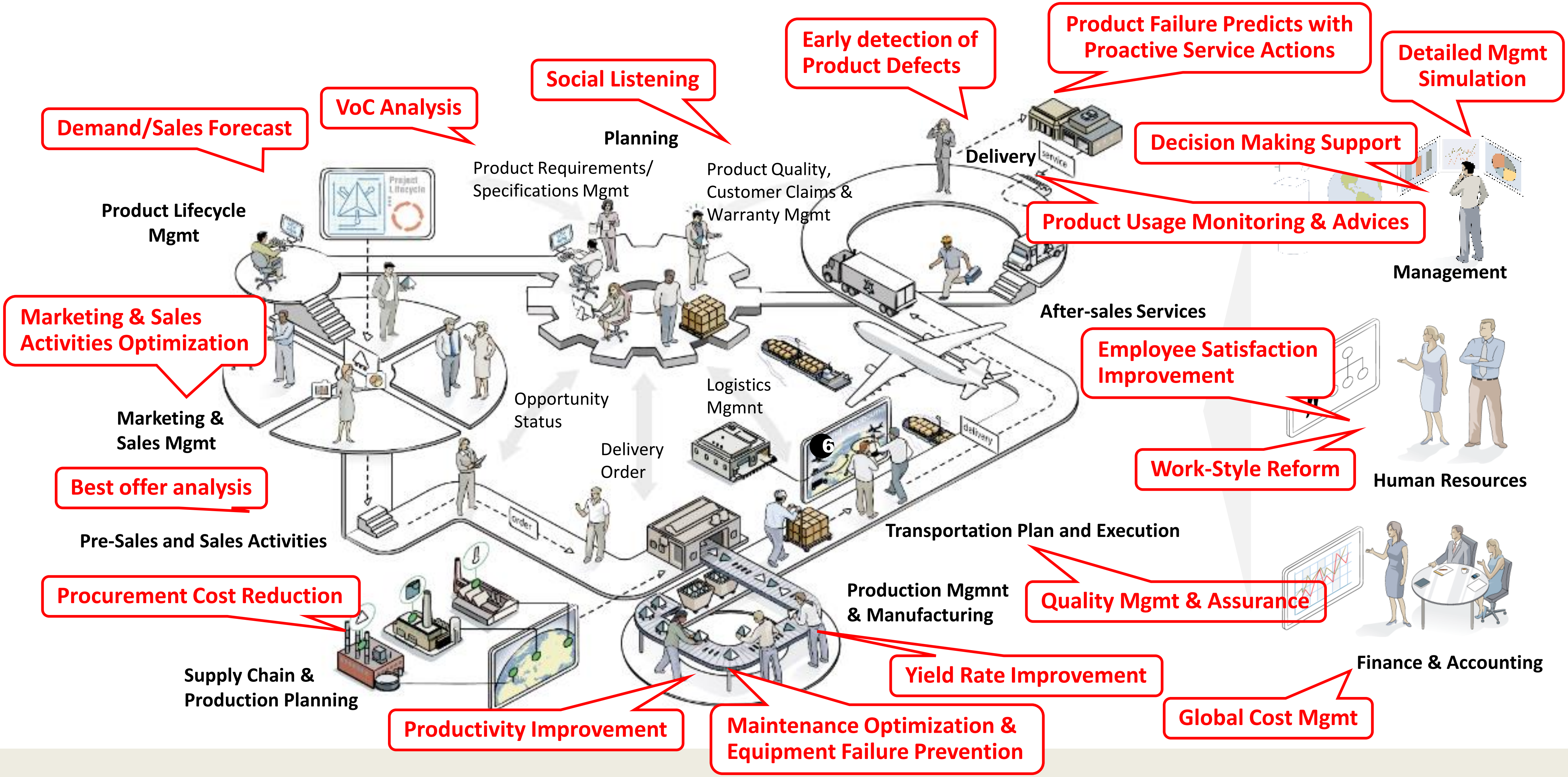
Copyright © 2018, Oracle and/or its affiliates. All rights reserved. |

Enterprise use cases of Machine Learning on Big Data

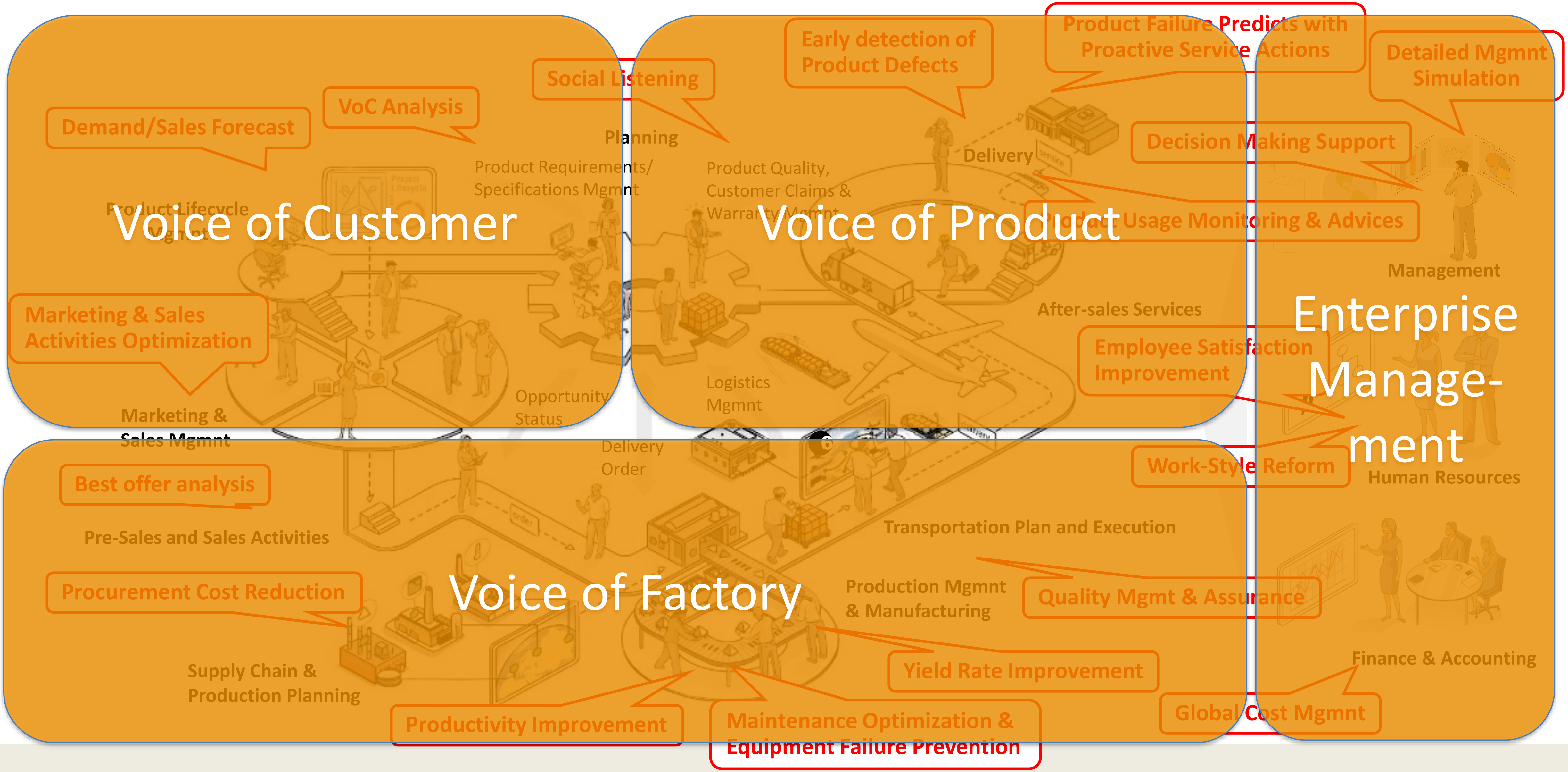
ORACLE

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. | Confidential – Oracle Internal/Restricted/Highly Restricted

“Machine Learning on Big Data” use case examples


Demand/Sales Forecast
VoC Analysis
Social Listening
Early detection of Product Defects
Product Failure Predicts with Proactive Service Actions
Detailed Mgmt Simulation
Decision Making Support
Product Usage Monitoring & Advices
Marketing & Sales Activities Optimization
Employee Satisfaction Improvement
Work-Style Reform
Best offer analysis
Procurement Cost Reduction
Quality Mgmt & Assurance
Yield Rate Improvement
Global Cost Mgmt
Productivity Improvement
Maintenance Optimization & Equipment Failure Prevention
Planning
Product Requirements/ Specifications Mgmt
Product Quality, Customer Claims & Warranty Mgmt
Delivery
service
Product Lifecycle Mgmt
Project Lifecycle
Management
After-sales Services
Marketing & Sales Mgmt
Opportunity Status
Logistics Mgmnt
Delivery Order
delivery
order
Human Resources
Pre-Sales and Sales Activities
Transportation Plan and Execution
Production Mgmnt & Manufacturing
Finance & Accounting
Supply Chain & Production Planning

“Machine Learning on Big Data” use case examples



Voice of Customer

Demand/Sales Forecast

VoC Analysis

Social Listening

Marketing & Sales Activities Optimization

Planning

Product Requirements/ Specifications Mgmnt

Product Lifecycle Mgmnt

Opportunity Status

Marketing & Sales Mgmnt

Voice of Product

Early detection of Product Defects

Product Failure Predicts with Proactive Service Actions

Product Usage Monitoring & Advices

Employee Satisfaction Improvement

Product Quality, Customer Claims & Warranty Mgmnt

Delivery

After-sales Services

Logistics Mgmnt

Enterprise Manage-ment

Detailed Mgmnt Simulation

Decision Making Support

Work-Style Reform

Global Cost Mgmnt

Management

Human Resources

Finance & Accounting

Voice of Factory

Best offer analysis

Procurement Cost Reduction

Quality Mgmt & Assurance

Yield Rate Improvement

Productivity Improvement

Maintenance Optimization & Equipment Failure Prevention

Delivery Order

Pre-Sales and Sales Activities

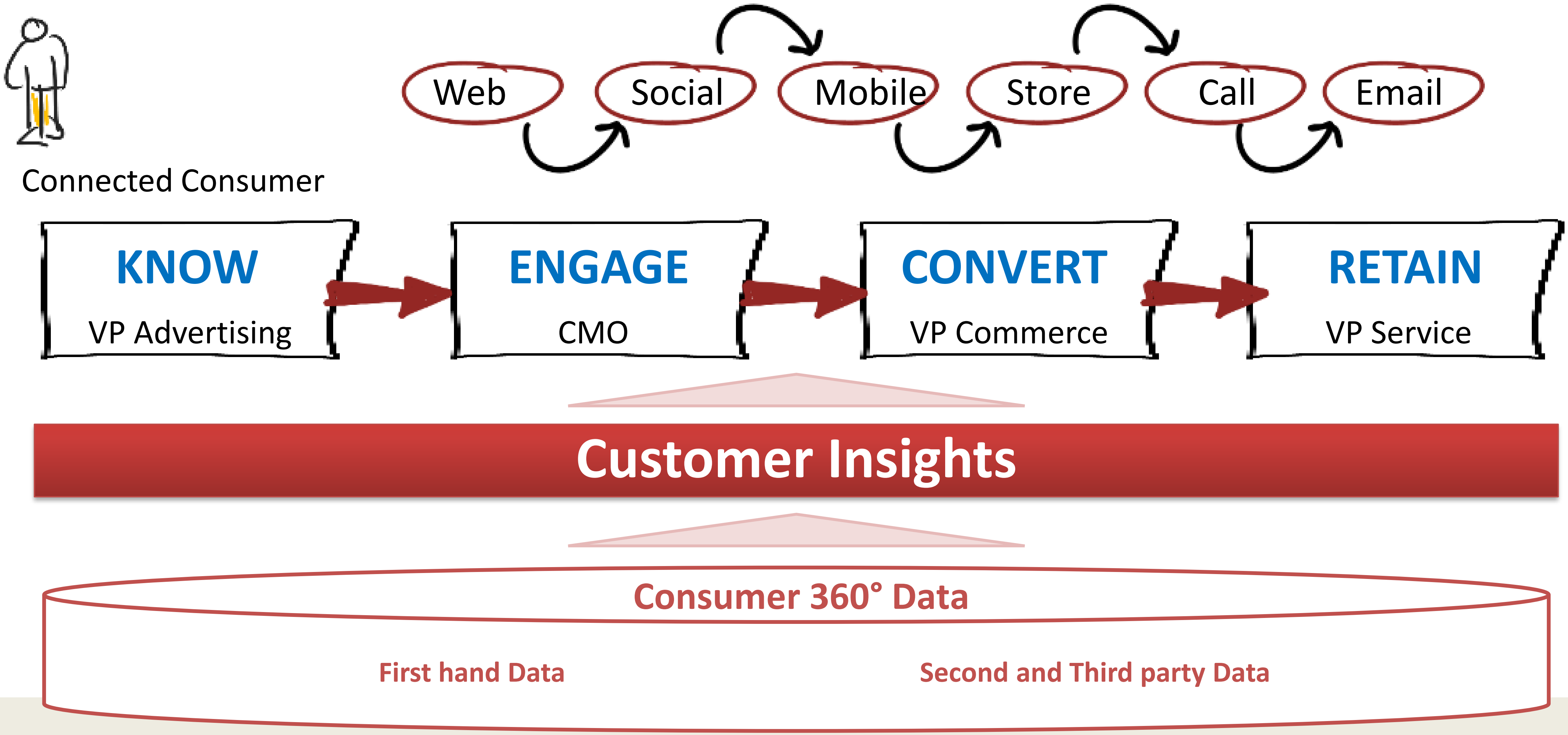
Transportation Plan and Execution

Production Mgmnt & Manufacturing

Supply Chain & Production Planning

Voice-of-Customer use case

Complete refoundation of customer interaction, thanks to knowledge and usage of all customer-related data

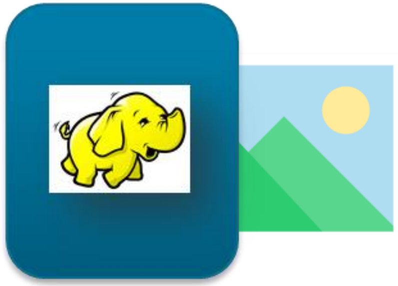


Considerations for a Successful Data Analytics Project

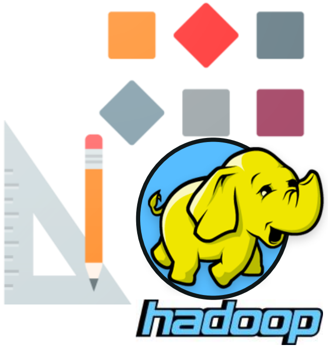
Copyright © 2017, Oracle and/or its affiliates. All rights reserved. | Confidential – Oracle Internal/Restricted/Highly Restricted

Implementing Big Data Projects: Overview

1. Information Data Management Architecture

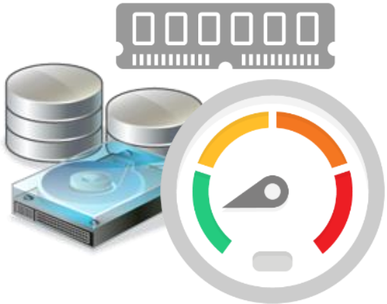


2. Understanding the Hadoop Ecosystem



3. Hadoop Architectural Patterns, General Rules, and Recommendations

4. Big Data Appliance Management Tools



5. Resource Management

6. File Types and Compression

7. Security

8. Back-up and Disaster Recovery

9. Data Integration Tools

10. End-user Tools

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Hadoop Ecosystem Projects

Hadoop Project	Type	Purpose
Hive	MR abstraction	Provide SQL-like (HiveQL) Functionality
Pig	MR abstraction	Provide functional programming interface
HBase	NoSQL database	Fast, scalable NoSQL engine
Hue	Web GUI	Web interface for end-users
Cloudera Manager	Web GUI for managing CDH	Web interface for administrators
Sqoop	Data import and export	Import and export data between RDBMS and HDFS
Flume	Data import	Stream real-time data into HDFS
Oozie	Workflow builder	Workflow scheduler
Impala	Run SQL queries	Run real-time SQL queries
Avro	Data interchange protocol	Data serialization and De-serialization
Mahout	Machine learning libraries	Algorithms and scripts
Kafka	Distributed streaming platform	Distributed service bus

ORACLE

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Hadoop: Use Cases and Data Generated

Types of Analyses that use Hadoop:

- *Market analysis*
- *Product recommendations*
- Demand forecasting
- *Fraud detection*
- Text mining
- *Index building*
- Graph creation and analysis
- Pattern recognition
- Collaborative filtering
- Prediction models
- Sentiment analysis
- Risk assessment

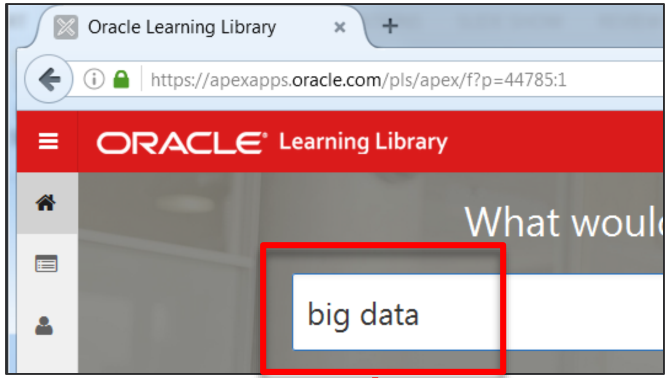
Types of data generated:

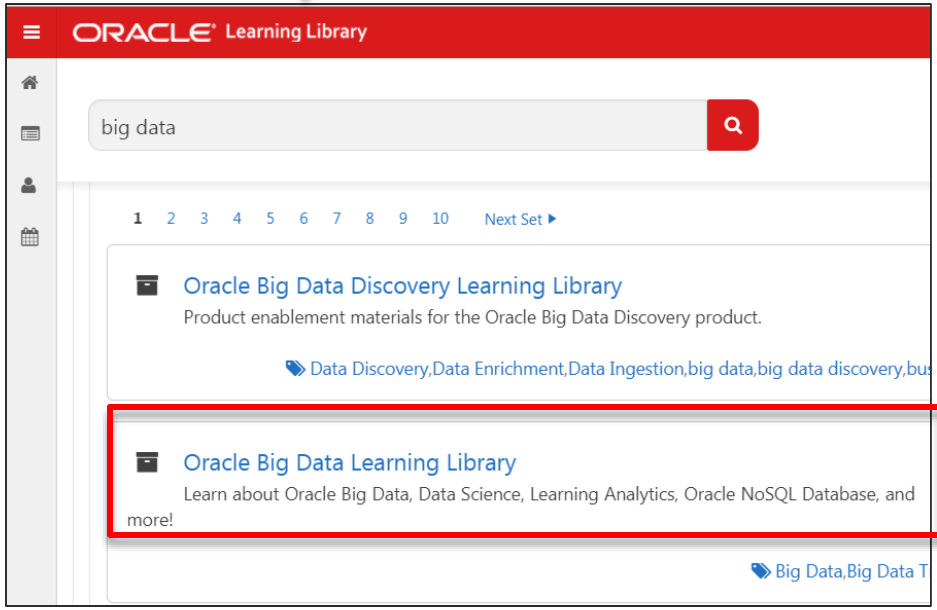
- Financial transactions
- Sensors data
- Server logs
- Analytics
- Email and text messages
- Social media

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Additional Resources: Oracle Learning Library (OLL)

http://www.oracle.com/goto/oll





Oracle Big Data Discovery Learning Library

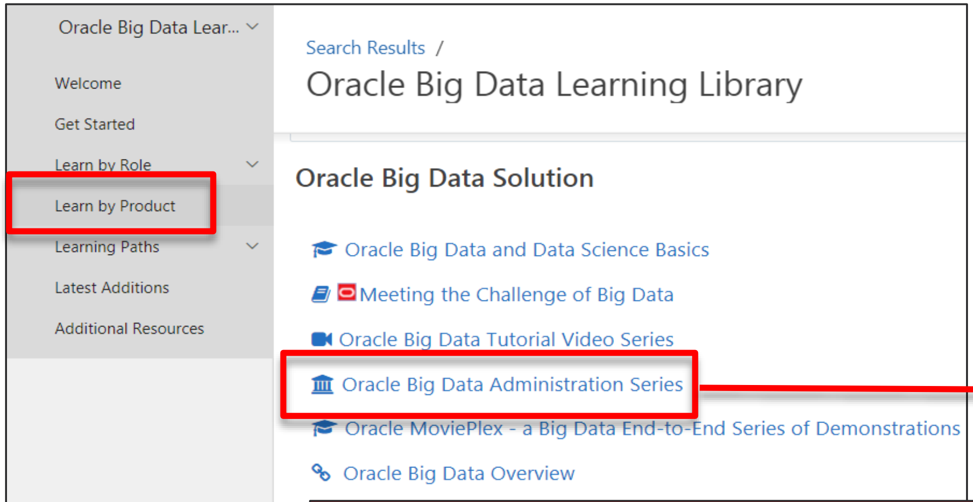
Product enablement materials for the Oracle Big Data Discovery product.

Data Discovery,Data Enrichment,Data Ingestion,big data,big data discovery,bu

Oracle Big Data Learning Library

Learn about Oracle Big Data, Data Science, Learning Analytics, Oracle NoSQL Database, and more!

Big Data,Big Data T

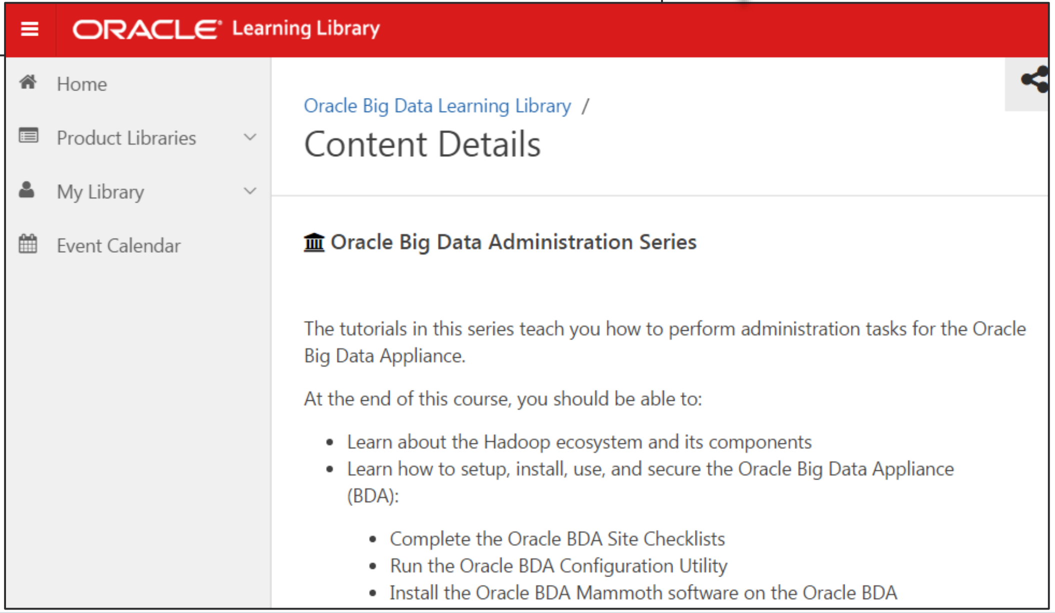


Search Results /

Oracle Big Data Learning Library

Oracle Big Data Solution

- Oracle Big Data and Data Science Basics
- Meeting the Challenge of Big Data
- Oracle Big Data Tutorial Video Series
- Oracle Big Data Administration Series
- Oracle MoviePlex - a Big Data End-to-End Series of Demonstrations
- Oracle Big Data Overview



Oracle Big Data Learning Library /

Content Details

Oracle Big Data Administration Series

The tutorials in this series teach you how to perform administration tasks for the Oracle Big Data Appliance.

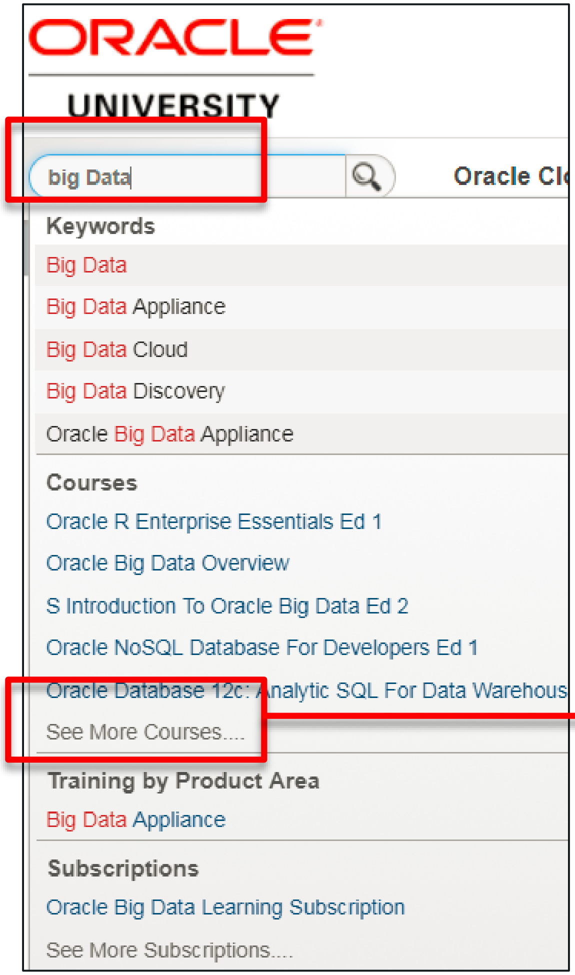
At the end of this course, you should be able to:

- Learn about the Hadoop ecosystem and its components
- Learn how to setup, install, use, and secure the Oracle Big Data Appliance (BDA):
  - Complete the Oracle BDA Site Checklists
  - Run the Oracle BDA Configuration Utility
  - Install the Oracle BDA Mammoth software on the Oracle BDA

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Oracle University Courses

education.oracle.com



Keywords

- Big Data
- Big Data Appliance
- Big Data Cloud
- Big Data Discovery
- Oracle Big Data Appliance

Courses

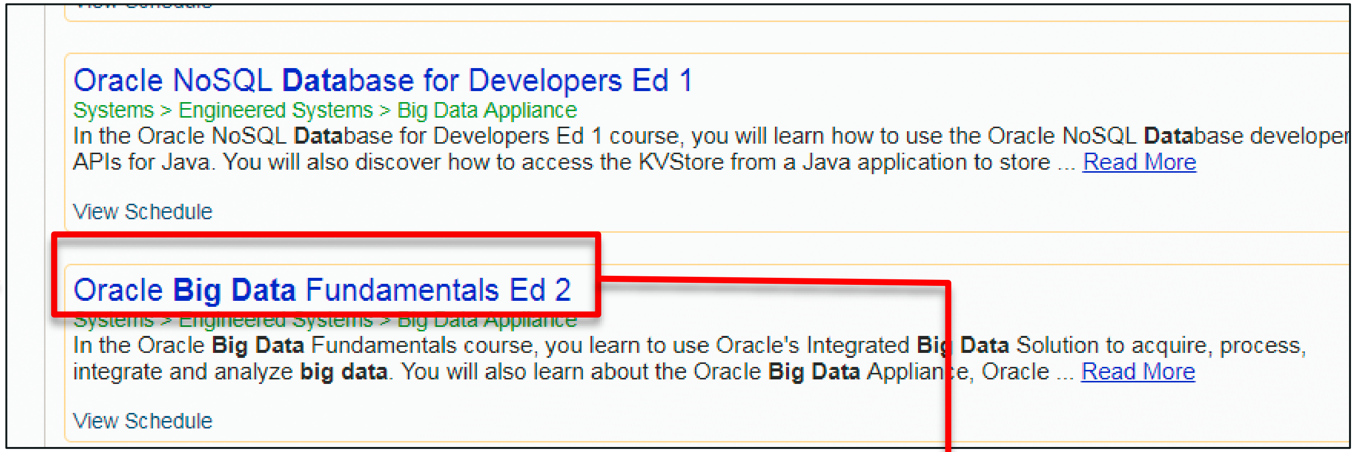
- Oracle R Enterprise Essentials Ed 1
- Oracle Big Data Overview
- S Introduction To Oracle Big Data Ed 2
- Oracle NoSQL Database For Developers Ed 1
- Oracle Database 12c: Analytic SQL For Data Warehous
- See More Courses....

Training by Product Area

- Big Data Appliance

Subscriptions

- Oracle Big Data Learning Subscription
- See More Subscriptions....



Oracle NoSQL **Data**base for Developers Ed 1

Systems > Engineered Systems > Big Data Appliance

In the Oracle NoSQL **Data**base for Developers Ed 1 course, you will learn how to use the Oracle NoSQL **Data**base developer APIs for Java. You will also discover how to access the KVStore from a Java application to store ... Read More

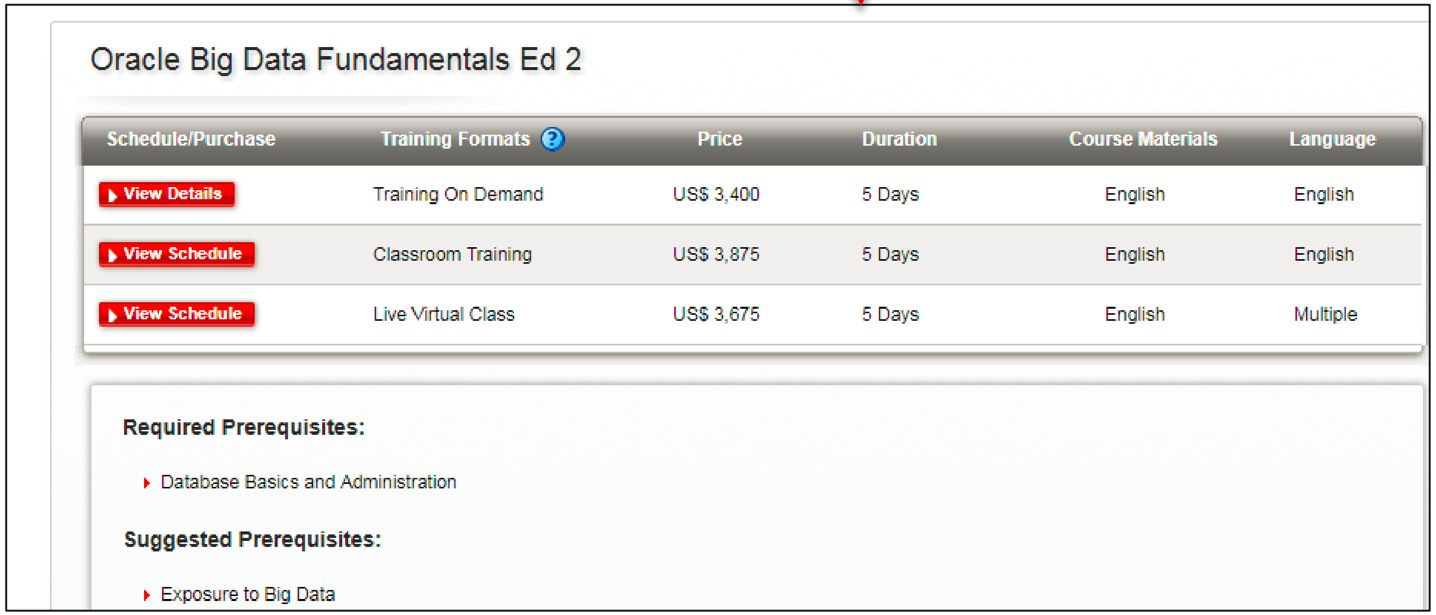
View Schedule

Oracle **Big Data** Fundamentals Ed 2

Systems > Engineered Systems > Big Data Appliance

In the Oracle **Big Data** Fundamentals course, you learn to use Oracle's Integrated **Big Data** Solution to acquire, process, integrate and analyze **big data**. You will also learn about the Oracle **Big Data** Appliance, Oracle ... Read More

View Schedule



Oracle Big Data Fundamentals Ed 2

Schedule/Purchase	Training Formats	Price	Duration	Course Materials	Language
▶ View Details	Training On Demand	US$ 3,400	5 Days	English	English
▶ View Schedule	Classroom Training	US$ 3,875	5 Days	English	English
▶ View Schedule	Live Virtual Class	US$ 3,675	5 Days	English	Multiple

Required Prerequisites:

- Database Basics and Administration

Suggested Prerequisites:

- Exposure to Big Data

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Resources

Topic	URL
Information Management and Big Data: A Reference Architecture	http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf
Architecting Big Data	https://www.youtube.com/watch?v=JT4qjEOU3KQ
Major goals of HDFS design	http://www.itversity.com/topic/major-goals-of-hdfs-design/
HDFS Design Concepts	http://hadooptutorial.info/hdfs-design-concepts/
NoSQL Databases	http://nosql-database.org/
Apache HBase Do’s and Don’ts	http://blog.cloudera.com/blog/2011/04/hbase-dos-and-donts/
HBase	https://www.slideshare.net/sawjd/h-base-20140613
Lambda Architecture	http://lambda-architecture.net/
Flafka: Apache Flume Meets Apache Kafka for Event Processing	http://blog.cloudera.com/blog/2014/11/flafka-apache-flume-meets-apache-kafka-for-event-processing/
Architectural Patterns for Near Real-Time Data Processing with Apache Hadoop	http://blog.cloudera.com/blog/2015/06/architectural-patterns-for-near-real-time-data-processing-with-apache-hadoop/
Sample small files	https://github.com/filanovskiy/catchSmallBlocks

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Resources

Topic	URL
Kerberos (protocol)	https://en.wikipedia.org/wiki/Kerberos_(protocol)
Instructions to Enable/Disable AD Kerberos on Oracle Big Data Appliance with Mammoth V4.2 Release (Doc ID 2029378.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=2029378.1
Instructions to Enable Kerberos on Oracle Big Data Appliance with Mammoth V3.1/V4.* Release (Doc ID 1919445.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=1919445.1
How to Set up a Cross-Realm Trust to Configure a BDA MIT Kerberos Enabled Cluster with Active Directory on BDA V4.5 and Higher (Doc ID 2198152.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=2198152.1
Understanding 'kinit' and Options for Authenticating All Nodes of a BDA Cluster (Doc ID 2004648.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=2004648.1
How to Enable/Disable HDFS Transparent Encryption on Oracle Big Data Appliance V4.4 with bdacli (Doc ID 2111343.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=2111343.1
How to Add or Remove Sentry on Oracle Big Data Appliance v4.2 or Higher with bdacli (Doc ID 2052733.1)	https://support.oracle.com/epmos/faces/DocumentDisplay?id=2052733.1
Apache Sentry	http://blog.cloudera.com/blog/2016/03/apache-sentry-is-now-a-top-level-project/

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Resources

Topic	URL
Dynamic Resource Pools	https://www.cloudera.com/documentation/enterprise/5-6-x/topics/cm_mc_resource_pools.html
Static Resource Pools	https://www.cloudera.com/documentation/enterprise/5-6-x/topics/cm_mc_service_pools.html
Apache Hadoop YARN: Avoiding 6 Time-Consuming "Gotchas"	http://blog.cloudera.com/blog/2014/04/apache-hadoop-yarn-avoiding-6-time-consuming-gotchas/
Untangling Apache Hadoop YARN, Part 1: Cluster and YARN Basics	https://blog.cloudera.com/blog/2015/09/untangling-apache-hadoop-yarn-part-1/
Untangling Apache Hadoop YARN, Part 2: Global Configuration Basics	http://blog.cloudera.com/blog/2015/10/untangling-apache-hadoop-yarn-part-2/
Untangling Apache Hadoop YARN, Part 3: Scheduler Concepts	http://blog.cloudera.com/blog/2016/01/untangling-apache-hadoop-yarn-part-3/
Untangling Apache Hadoop YARN, Part 4: Fair Scheduler Queue Basics	http://blog.cloudera.com/blog/2016/06/untangling-apache-hadoop-yarn-part-4-fair-scheduler-queue-basics/
Untangling Apache Hadoop YARN, Part 5: Using FairScheduler queue properties	https://blog.cloudera.com/blog/2017/02/untangling-apache-hadoop-yarn-part-5-using-fairscheduler-queue-properties/

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |

Resources

Topic	URL
Hadoop Compression. Compression rate – Part1	https://blogs.oracle.com/datawarehousing/hadoop-compression-compression-rate-part1
Hadoop Compression. Choosing compression codec – Part2	https://blogs.oracle.com/datawarehousing/hadoop-compression-choosing-compression-codec-part2
Secure your Hadoop Cluster	https://blogs.oracle.com/datawarehousing/secure-your-hadoop-cluster

Copyright © 2017, Oracle and/or its affiliates. All rights reserved. |