Big Data Engineering In details

From Beginner to Professional

Mostafa Alaa Mohamed
Senior Big Data Engineer
MoustafaAlaa Moustafa Alaa @Moustafa_alaa22
mustafa.alaa.mohamed@gmail.com

[1]Big Data & Analytics Department, Epam Systems

The Definitive Guide to Big Data Engineering Tasks

Table of Contents I

Table of Contents II

Table of Contents III

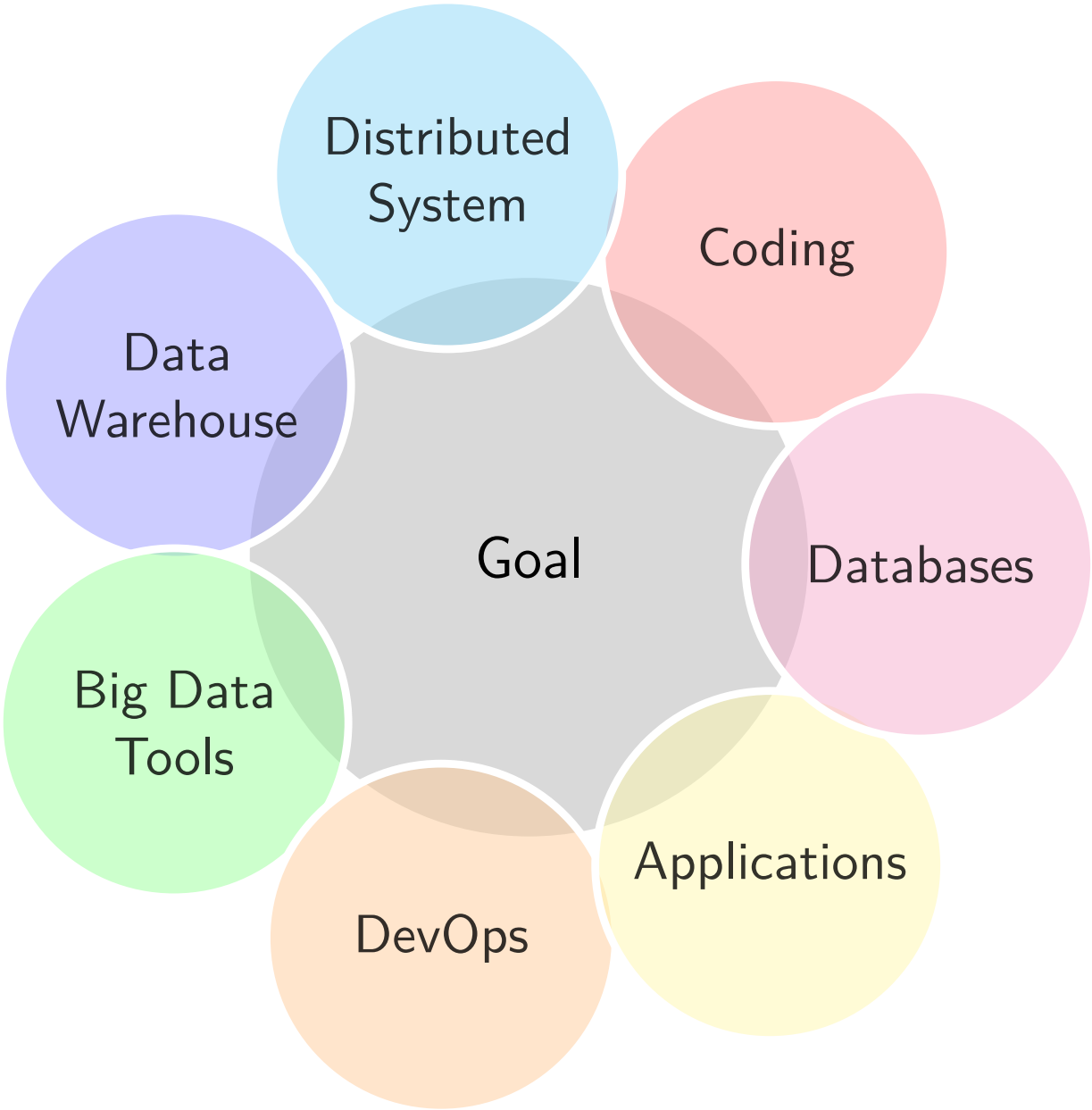
Table of Contents IV

Table of Contents V

Table of Contents VI

Course Introduction

Course Target



Learning Objectives

- Understand the data management life-cycle.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.
- Applying machine learning over Big Data.

Learning Objectives

- Understand the data management life-cycle.
- Illustrate the basics of distributed systems concepts
- Be familiar with ETL for (Batch/Steaming) data over distributed systems ex: Hadoop & Spark.
- Apply QA and testing for the data pipeline cycle.
- Automate the Data life-cycle process End-to-End.
- Building real-life examples.
- Applying machine learning over Big Data.
- Understanding of the DevOps tools and functions in data life-cycle.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.
- Ask your questions.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.
- Ask your questions.
- Join the online meeting or discussions.

Getting max benefit from this course

Take the course advantage

- Follow the videos order as described.
- Read the references for each section (including the implementation of the examples if exists).
- Repeat the lecture code with your own.
- Do the assignments.
- Ask your questions.
- Join the online meeting or discussions.

Assignments and Labs

Remark

- Full project code.

Assignments and Labs

Remark

- Full project code.
- Notebooks (Jupyter or Zeppelin).

Assignments and Labs

Remark

- Full project code.
- Notebooks (Jupyter or Zeppelin).
- Read the reference.

Textbooks-1

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.

Textbooks-1

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau

Textbooks-1

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.

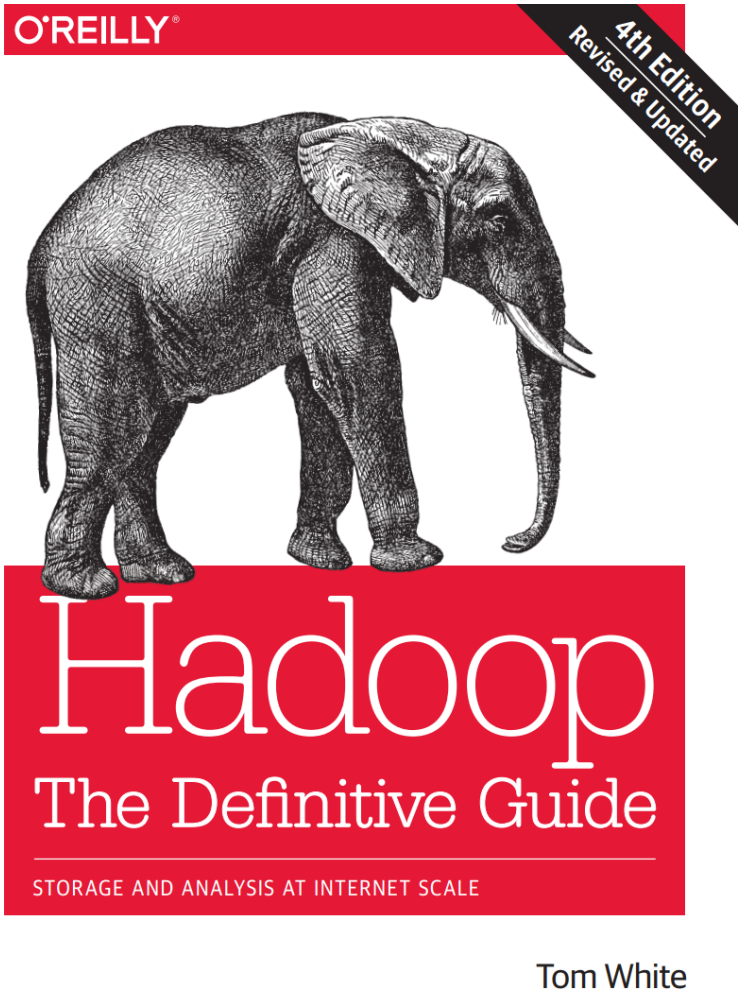
Textbooks-1

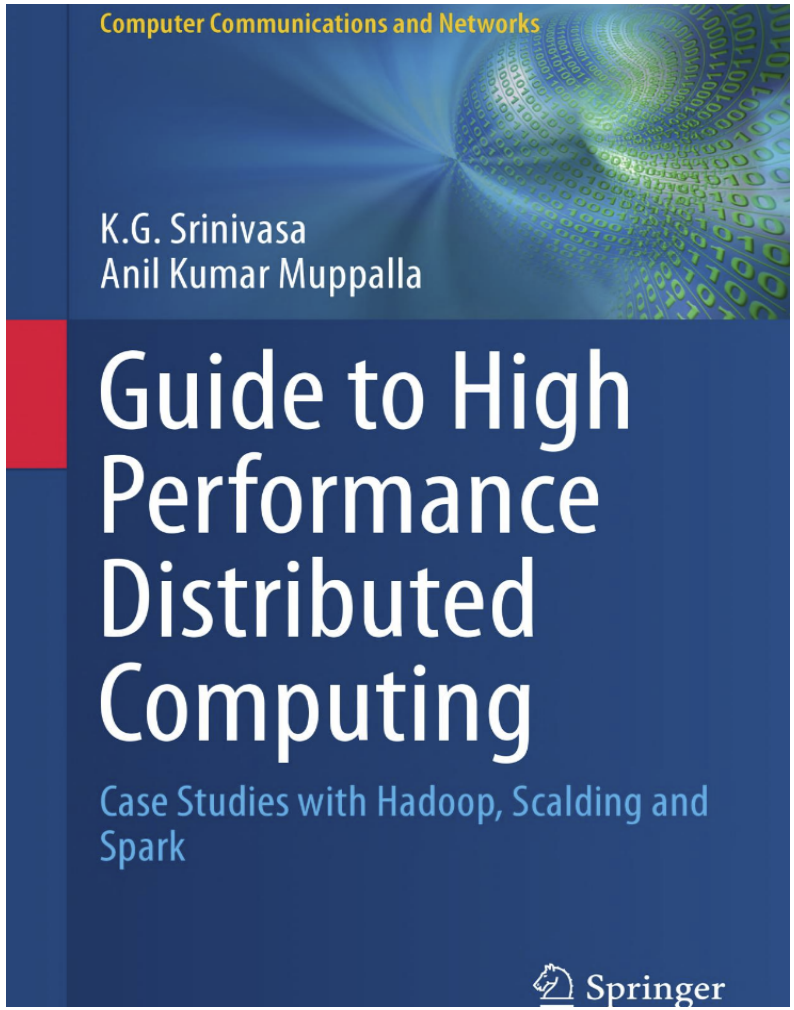
- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.
- Kafka: The Definitive Guide by Todd Palino, Gwen Shapira, Neha Narkhede.

Textbooks-1

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale 4th Edition by Tom White.
- Learning Spark by Matei Zaharia, Patrick Wendell, Andy Konwinski, Holden Karau
- High Performance Spark Best Practices for Scaling and Optimizing Apache Spark By Holden Karau, Rachel Warren.
- Kafka: The Definitive Guide by Todd Palino, Gwen Shapira, Neha Narkhede.
- Guide to High Performance Distributed Computing: Case Studies with Hadoop, Scalding and Spark (Computer Communications and Networks) 2015th Edition

Textbooks-2

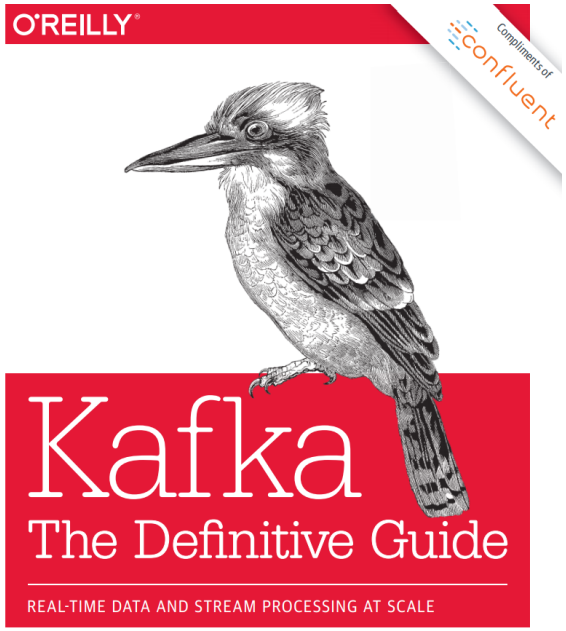

O'REILLY
4th Edition
Revised & Updated
Hadoop
The Definitive Guide
STORAGE AND ANALYSIS AT INTERNET SCALE
Tom White


Computer Communications and Networks
K.G. Srinivasa
Anil Kumar Muppalla
Guide to High Performance Distributed Computing
Case Studies with Hadoop, Scalding and Spark
Springer

Textbooks-3

Holden Karau, Andy Konwinski,
Patrick Wendell & Matei Zaharia

Holden Karau &
Rachel Warren



Neha Narkhede,
Gwen Shapira & Todd Palino

Ugly but important

- User stories or technical discussions are not related to any of my current work or my previous companies.
- I am working at EPAM Systems. My company approved me for doing this online course public but the materials are not reviewed or assessed by my company. It is on my own responsibilities.

Introduction To Distributed Systems (Hadoop as as example)

Chapter Objectives

- What is data management?

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.

Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.

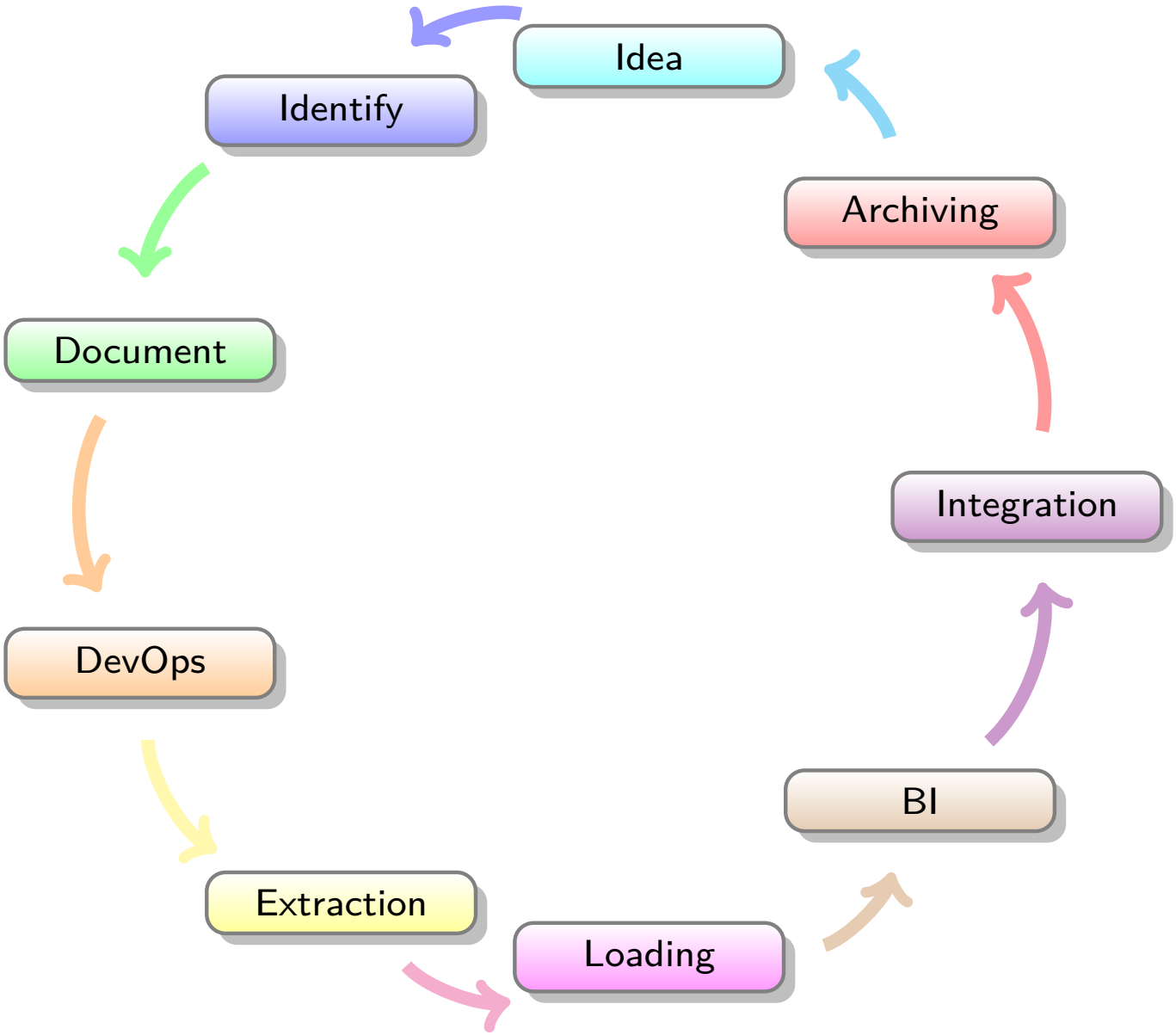
Chapter Objectives

- What is data management?
- Introduction to distributed systems concepts
- Why we need Hadoop?
- Understand the concept of HDFS and Map-Reduce.
- Developing Map-Reduce applications.
- Using Hive QL over Map-Reduce.
- Hadoop advantages and disadvantages with use cases?

Data Management

- Data are a product.
- Data product has a life-cycle as following (simplified):
  - **Question**, Idea, or service.
  - **Identifying** the source of information and the data type ex: (text, images, videos, audio, or sensors).
  - **Document** all details regarding the data including quality, security, efficiency, and access (consideration during the cycle).
  - Delivery automation (Tools and Process) AKA **DevOps** cycle.
  - **Extraction** Process (collection).
  - **Transformation** ex: (cleansing, Apply business logic, Organize).
  - **Loading** or store the transformed data based on our usage or use case.
  - Business Intelligence (**BI**) or data discovery (continues process).
  - **Integration** and publishing.
  - Data retention or **archiving** process ex: (Hot or Cold storage).

Data Management Life-Cycle



Motivation to Data Warehouse

- Data could be a product for some companies.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
  - Performance.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
  - Performance.
  - Integration.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
  - Performance.
  - Integration.
  - Applying analytical functions.

Motivation to Data Warehouse

- Data could be a product for some companies.
- It could be decision support for other products or businesses.
- Reporting the results after pass the data life-cycle will be from storage (Database).
- There are some challenges facing the people who work on data management backend:
  - Performance.
  - Integration.
  - Applying analytical functions.
- Vendors who are working to solve the above challenges creating their own product of DWH and their ultimate work is to optimize the above points.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.
- Most of DWH can't solve the online transactions similar to the transaction DB.

DWH vs Operational databases

- Operational databases (Transactions DB) still working as the backend for the products.
- Data warehouse mainly works as centralized storage for all the source systems regardless of the product type or their functionality.
- Data warehouse designed to solve the **huge amount of data**.
- Most of DWH can't solve the online transactions similar to the transaction DB.
- Transactions databases have a performance issue while handling a huge amount of data. So, analysis of a huge amount of data (including historical data) we used DWH for this purpose. On the other hand Transactions DB used for online or short historical data based on product type and requirements.

DWH vs Operational databases

#	Transactions DB	DW
Volume	GB/TB	TB/
Historical	Short-term	Long-
rows	<1000M	1000
Orientation	Product	Subject or m
Business Units	Product team	Multi organiz
Normalization	Normalized	Not required (De-normal
Data Model	Relational	Star Schema
Intelligence	Reporting	Advanced reporting a
Use cases	Online transactions & operations	Centeralized s

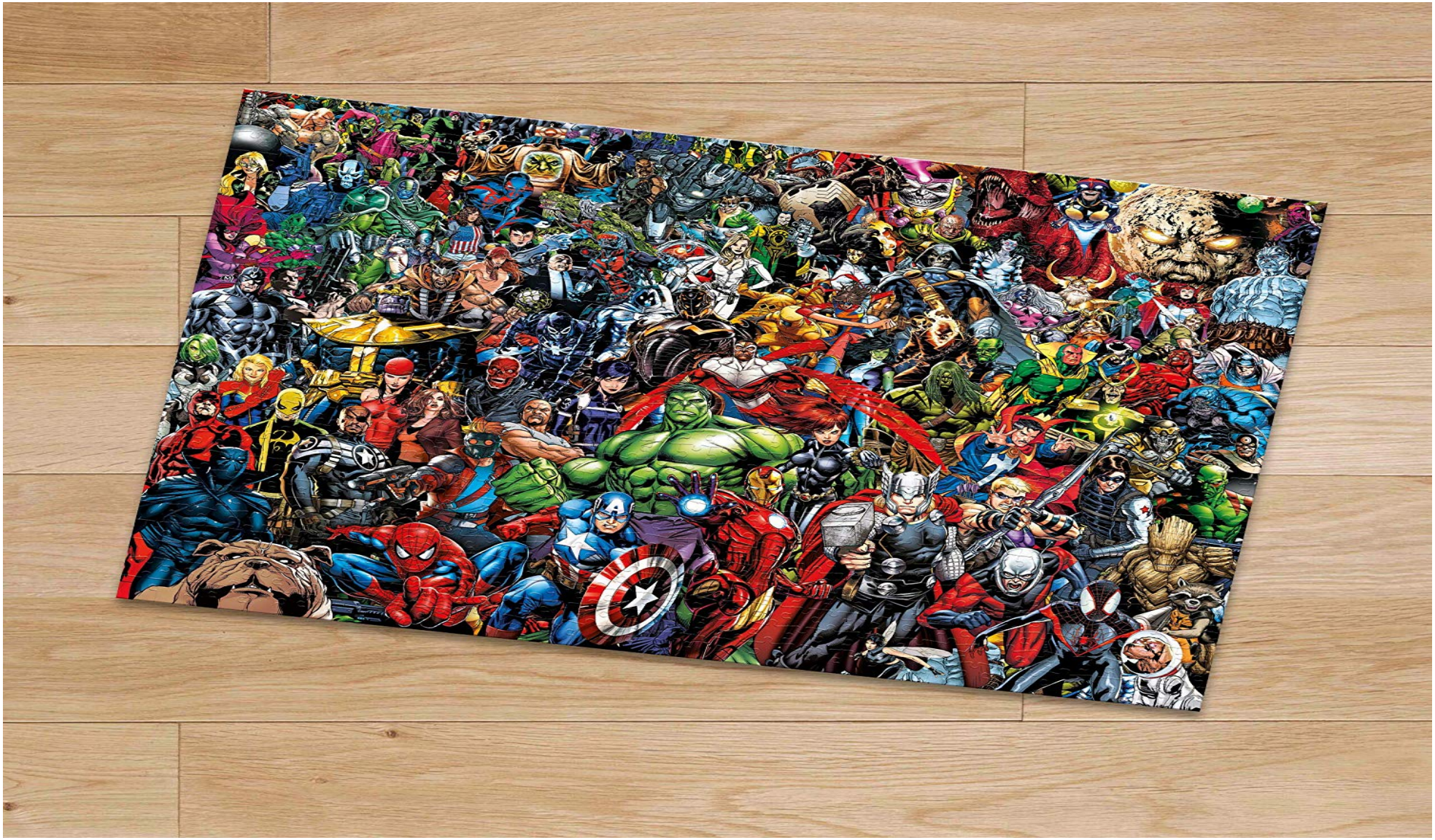
Table: Data Representation Combination Matrix

Transnational DB Use cases

Transnational DB Use cases

DWH Use cases

DWH Use cases



DWH Use cases

User stories Telecom company.
It has a CRM System backend database reporting the sales. vs Another backend database contains the CRM, Telecom signaling data, IN charging system, Billing
Decision is related to sales or CRM. Decision is related to company strategies.
Analytical model checking the fraud which require a CRM data with customer locations from signaling with Billing details from CAR table.
managing risk of the project in Transaction vs DWH
data model comparison

Cold storage vs Hot storage

some details about hot vs cold storage,

DWH Characteristics

some details about hot vs cold storage,

Distributed Systems Concepts

- Any Big Data solution working based distributed systems.

Distributed Systems Concepts

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Hadoop Architecture

- Any Big Data solution working based distributed systems.

Hadoop Architecture

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Storage

- Any Big Data solution working based distributed systems.

Storage

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

YARN

- Any Big Data solution working based distributed systems.

YARN

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Hadoop I/O

- Any Big Data solution working based distributed systems.

Hadoop I/O

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Processing

- Any Big Data solution working based distributed systems.

Processing

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Map-Reduce

- Any Big Data solution working based distributed systems.

Map-Reduce

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Map-Reduce Components

- Any Big Data solution working based distributed systems.

Map-Reduce Components

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Word-Count Example

- Any Big Data solution working based distributed systems.

Word-Count Example

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Hive

- Any Big Data solution working based distributed systems.

Hive

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Function Programming

Spark Framework

Spark Framework: Spark Philosophy towards the Engine and the Programming languages

- Any Big Data solution working based distributed systems.

Spark Framework: Spark Philosophy towards the Engine and the Programming languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Framework: Spark Basics

- Any Big Data solution working based distributed systems.

Spark Framework: Spark Basics

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Basics

- Any Big Data solution working based distributed systems.

Spark Basics

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.

Spark Programming using RDDs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.

Spark Datasets/Dataframe

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark on Production

- Any Big Data solution working based distributed systems.

Spark on Production

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark on Production

- Any Big Data solution working based distributed systems.

Spark on Production

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Batch Processing

- Any Big Data solution working based distributed systems.

Spark For Batch Processing

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Building custom input and output connector using Spark

- Any Big Data solution working based distributed systems.

Building custom input and output connector using Spark

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Streaming

- Any Big Data solution working based distributed systems.

Spark Streaming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.

Spark using other Programming Languages

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark For Data Scientist

- Any Big Data solution working based distributed systems.

Spark For Data Scientist

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.

Spark Graph Dataframe/Graphx

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.

Tuning your Spark Jobs

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Real World Applications

Massaging Systems

Elastic

NOSQL

Data Orchestration

Appendix

Appendix A- Shell Programming

- Any Big Data solution working based distributed systems.

Appendix A- Shell Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix B- Java Programming

- Any Big Data solution working based distributed systems.

Appendix B- Java Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix C- Scala Programming

- Any Big Data solution working based distributed systems.

Appendix C- Scala Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix D- SQL Programming

- Any Big Data solution working based distributed systems.

Appendix D- SQL Programming

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix E- Oozie Orchestration

- Any Big Data solution working based distributed systems.

Appendix E- Oozie Orchestration

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix F- DWH Concepts and Data Modeling Design

- Any Big Data solution working based distributed systems.

Appendix F- DWH Concepts and Data Modeling Design

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix G- Machine Learning Concepts Data Engineers

- Any Big Data solution working based distributed systems.

Appendix G- Machine Learning Concepts Data Engineers

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?

Appendix H- Docker for Data Engineers

- Any Big Data solution working based distributed systems.

Appendix H- Docker for Data Engineers

- Any Big Data solution working based distributed systems.
- What is distributed systems in brief?