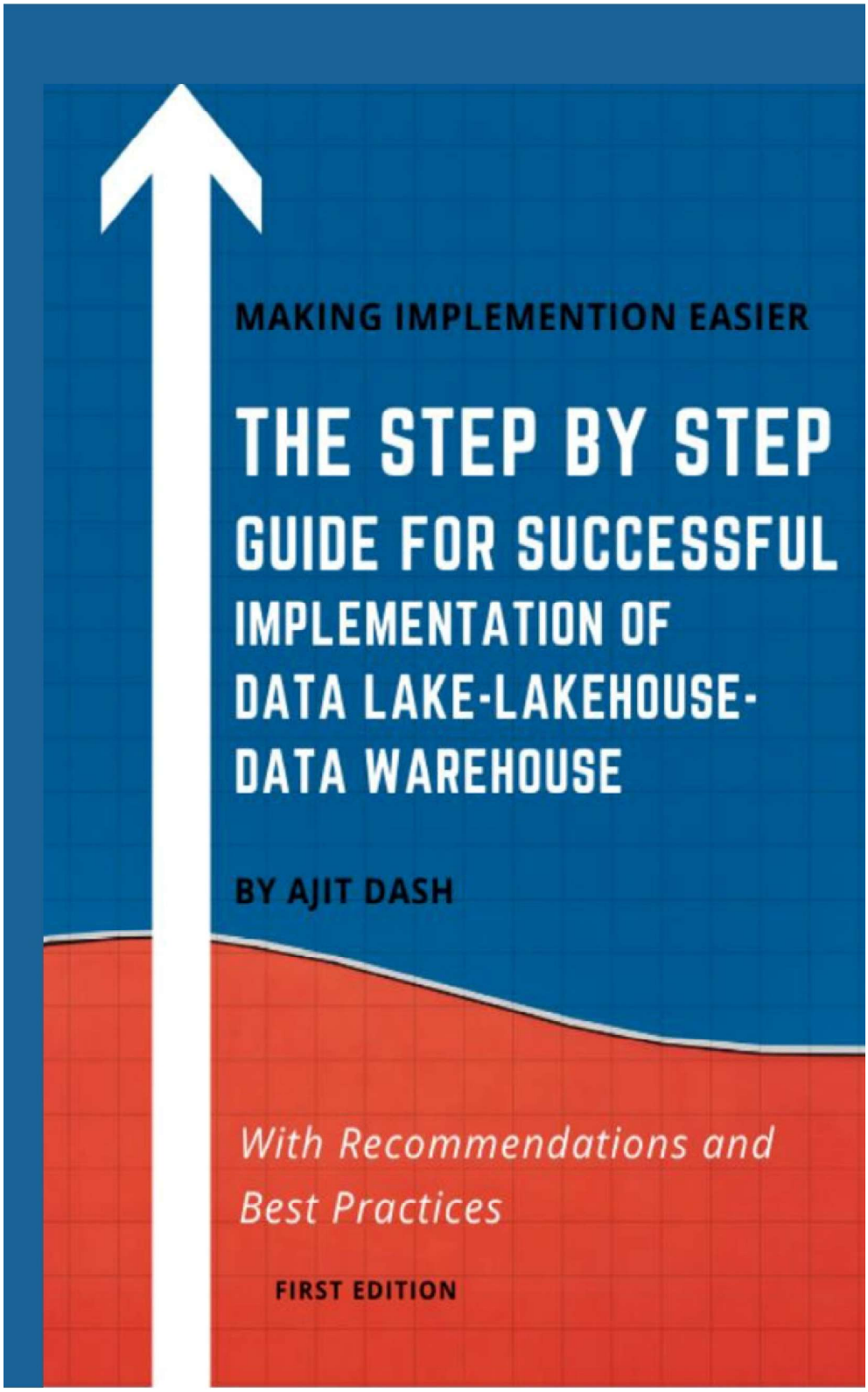

MAKING IMPLEMENTION EASIER
THE STEP BY STEP GUIDE FOR SUCCESSFUL IMPLEMENTATION OF DATA LAKE-LAKEHOUSE-DATA WAREHOUSE
BY AJIT DASH
With Recommendations and Best Practices
FIRST EDITION

THE STEP BY STEP GUIDE FOR SUCCESSFUL IMPLEMENTATION OF DATA LAKE-LAKEHOUSE-DATA WAREHOUSE

BY AJIT DASH

Dedicated to Almighty

ISBN 978-1-008-93911-0

All the views are given is sole the view of the writer.

Table of Contents

PREFACE

This book gives information about the Data Lake, Lakehouse and Delta Lake, best practices, and precautions for successful implementation. Our general understanding is that the Data Lake and Lakehouse is a storage place. While that is true, it also the brain of an enterprise; it must be designed as per the business need with proper intelligence in place. In here, a comparison of a Data Lake, Lakehouse, Delta Lake, and Data Warehouse are analyzed thoroughly. How a Lakehouse could be easily expanded to a Delta Lake is summarized. A step-by-step analysis, design, development, and implementation plans with all the challenges and best practices are described in detail.

Also, this book provides recommendations, best practices on various topics such as data processing, storage layers, downstream process, planning, data quality, data governance, data retention-backup, data visibility, security, automation, tools selection and physical/ logical architecture samples, etc.

WHOM IS THIS BOOK FOR?

As data becoming a crucial asset for any type of organization and it is playing a bigger role starting from decision making to creating innovations to meet the present and future goals. Organizations are realizing that by using data and data analytics they can overcome any obstacles or challenges coming their way. In a corporation a wrong analysis of data could cause havoc in the decision-making process (financial or deliverables).

Most companies are relying on past and present data and using this information companies are creating analytics that eventually help them in the right direction.

"THE STEP BY STEP GUIDE FOR SUCCESSFUL IMPLEMENTATION OF DATA LAKE - LAKEHOUSE -DATA WAREHOUSE " gives information about how the present and history data could be saved to get the right data insights to help the decision making process to meet a company's end goal.

This book is written with a practical approach on how to store the data, process the data, and deliver the data along with best practices.

This book will give knowledge about "DATA LAKE - LAKEHOUSE -DATA WAREHOUSE " and how to implement it as well as the best practices to ensure that the implementation goes right.

This could be helpful for technical as well as non-technical readers who want to know more about the “DATA LAKE - LAKEHOUSE -DATA WAREHOUSE " how to store the data and process it.

I recommend this book for business executives, technical executives, technologist, solution architecture, software developers, data developers, cloud practitioner and implantation specialist

CHAPTER 1

Understanding Datalake, Lakehouse and Data Warehouse

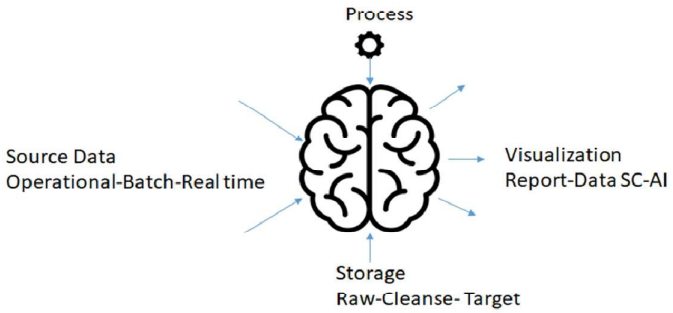
Data in an Organization

Business analytics and data insights are crucial in an organization to run the business and decision-making process efficiently. To achieve better analytics, business consumes data from different operational, application and third-party sources. Then stores these data to a destination after going through a cleansing and aggregation process known as ETL (Extract, Transfer, Load). These destinations could be a Data warehouse, Data Lake, Lakehouse, or Delta Lake. Data from the data sources could be structured, unstructured, or semi-structured. Also, the data frequency could be a stream, batch mode, or a mix of both batch and real-time.

Efficient storage and extractions are vital because it leads to better analytics. Let's discuss in detail how and where to store these source data and successfully process them.

What Data Lake Is

I like to compare a Data Lake with a human brain; the brain receives input information from different sources and processes it as needed, stores it for short or long-term usages, and provides the output. Furthermore, stored information could be leveraged for the required purpose in the future.



The Data Lake receives data from different data sources. The received data could be real-time, batch, or a mix of both batch and real-time.

The Data could be in any format - structured, unstructured, or semi-structured. The Data is received either by a push (i.e., data push to a storage layer by source system using extraction tool) or a pull (i.e., data pulled from the source layer to the storage layer) process in raw format to a Landing (storage) layer. From the Landing layer, it moves to the Staging layer as-it-is format (preferably using ELT - Extraction Load and Transfer process) and from here, it moves to the Cleanse layer where business rules apply before it moves to the Integration layer or a final Target layer. Further for downstream consumption, it moves data

from the Target layer to the Visibility layer, which may consist of a report, data, AI, or a data dump to a bucket with various formats (e.g. .TXT, XLS, CSV, JSON,XML etc.)

Data Lake uses a two-tier architecture:

- ✓ ETL / ELT from Source to Raw layer of the Data Lake
- ✓ Use ETL to move the data from the raw layer for further downstream consumption, such as moving the data to a data warehouse, reporting, or data analytics purpose

Advantages:

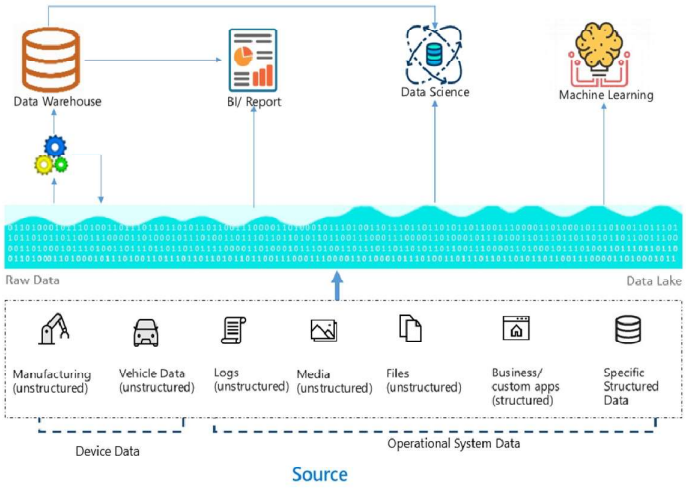
- ✓ Data from multiple source system reside in one place
- ✓ Data is always available from multiple sources. This makes it easy to slice and dice the data.
- ✓ Storage and compute are different makes processing faster
- ✓ Good for append only data

Disadvantages:

- ✓ Due to multiple ETL tiers, this makes the process complex
- ✓ High maintenance
- ✓ Data staleness is the major concern

- ✓ Stream and Batch data stored in one place, but extraction required extra work
- ✓ Multiple source data types and formats may bring complexity
- ✓ Data governance and data quality could be a challenge due to multiple variety and velocity of source data
- ✓ Data update can be very difficult it is good for append only data

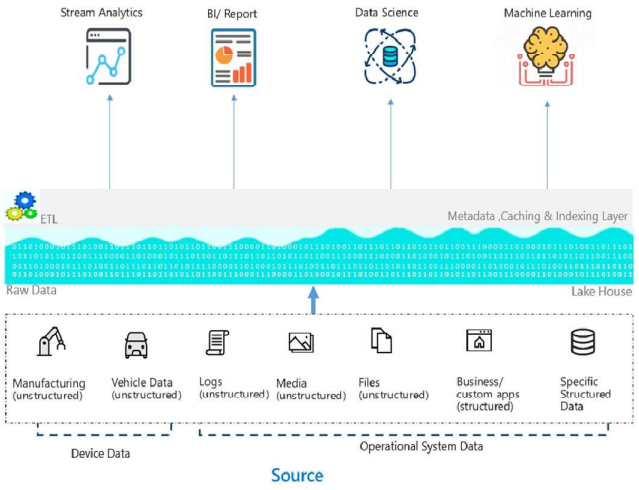
Data Lake



What Lakehouse Is

Lakehouse is a unique concept of storing all types of data in the same storage repository (Structured, Unstructured, or Semi-structured) as it is received in a common format such as Parquet. Simply, it's a combination of Data Lake and Data Warehouse functionalities. It provides scalability, flexibility and increases data consumption efficiency as compared to a Data Lake. Stored data could be used directly for ML, BI, or Data Warehouse. A Lakehouse resolves the challenges and limitations that arise in a Data lake and Data warehouse. It is an open architecture. For a Lakehouse you can use a custom ETL tool for data curation. Some of the software vendors like Snowflake or Databrick facilitate the creation of Lakehouse.

Lakehouse



Advantages: 

- ✓ It supports open architecture with structured, semi-structured, and unstructured data
- ✓ Allows ACID transaction, schema enforcement and evolution, supports star and snowflake
- ✓ BI reporting available directly from source data
- ✓ Data storage and computes separated which make data processing faster
- ✓ Machine learning, Data Sc, and BI data available from the same repository
- ✓ Storage data file format usually used Parquet
- ✓ It supports functionalities of both Data warehouse and Data Lake

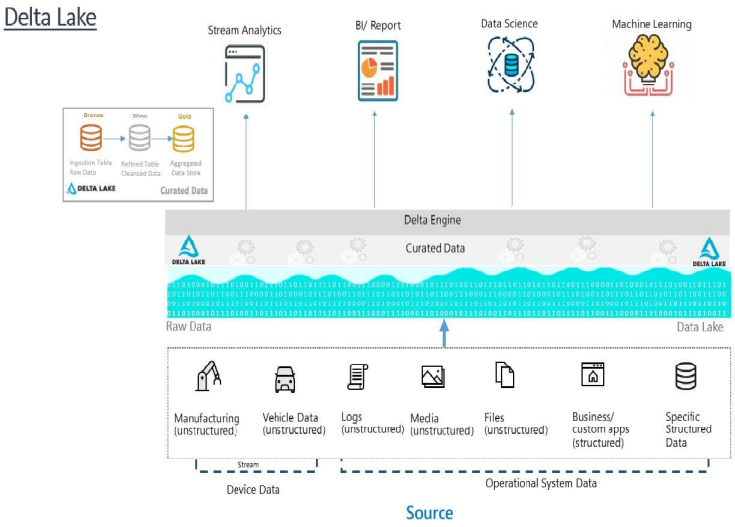
Disadvantages: 

- ✓ Data activities require well planning
- ✓ A powerful compute engine recommended for processing
- ✓ Data curation requires well planning

What Delta Lake Is:

Delta Lake is an open source project and it's a Lakehouse architecture on top of the Data Lake. Delta Lake stores structured, unstructured and semi-structured (batch, stream, or

combination) data in a common storage. All data files are commonly stored in a Parquet format. Delta Lake curates the Raw data (known as Bronze) then applies business rules before it moves to cleanse (known as Silver) and aggregates the data as required to make it business-ready (known as Gold). Gold data is in an aggregated format such as Datamart. Delta Lake uses Delta engine which is used for query and it is an Apache Spark compatible engine. Stream and Batch data could be read and queried for BI, ML, or Data Science. Delta Lake provides ACID transaction capability. As well, provides better data integrity, a single source of truth, and data governance. Delta Lake In other words is a combination of the Data Warehouse (where you can have a better governance) and Lakehouse (where you can process all kinds of source data)



Advantages: 

- ✓ Supports all data types structured, unstructured, and semi-structured
- ✓ ACID transaction available
- ✓ Batch and Stream data analysis possible
- ✓ Use Delta engine for reading the source data which is Spark compatible
- ✓ All types of data reside in a common layer and uses single access for data analysis
- ✓ Good data governance and data integrity
- ✓ Schema enforcement avoid automatically insert of the bad records during ingestion
- ✓ Provides Time travel which is an act as a check point for the rollback.
- ✓ Supports update, delete and merge which helps in change-data-capture, slowly-changing-dimension and Upserts streaming etc.

Disadvantages: 

- ✓ Performance measures depend on the compute node
- ✓ High power compute node usually recommended
- ✓ Suitable where batch and stream data operation needed

What Data Warehouse Is:

Data Warehouse is a repository for processed data. It uses the ETL tool to collect the raw data from multiple source systems such as CRM, accounting and operational systems then cleanse it after using the business rules and finally move to the data warehouse. A data warehouse could be relational, hybrid, or dimensional based on business requirements.

A Data Warehouse get benefited from the ACID compatibility but it not necessarily a requirement

Advantages:  TIP

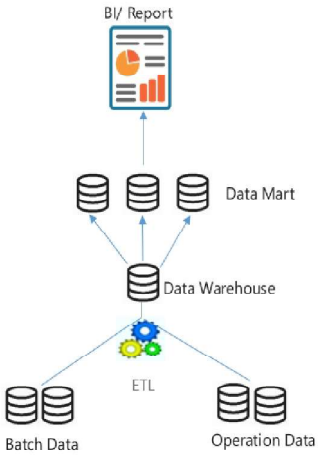
- ✓ Data granularity could be available as required by the business
- ✓ A data warehouse further extends into a data mart based on the business subject area requirements.
- ✓ Good for batch data storage
- ✓ Easy to implement Data Governance in a Data Warehouse
- ✓ Data Warehouse uses various types of ETL tools available in the market

Disadvantages: 

- ✓ Good for batch mode data or operational data (semi real time)

- ✓ Consumptions of stream data required further processing before moving to the Data Warehouse and it's more complex and not efficient
- ✓ ETL tool license fee is very expensive

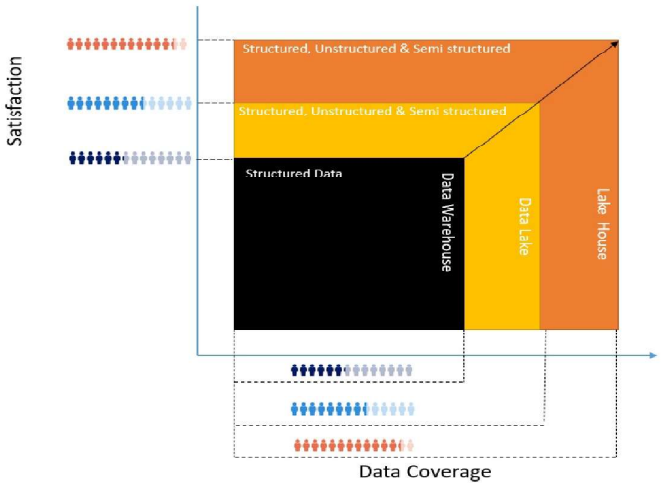
Data Warehouse

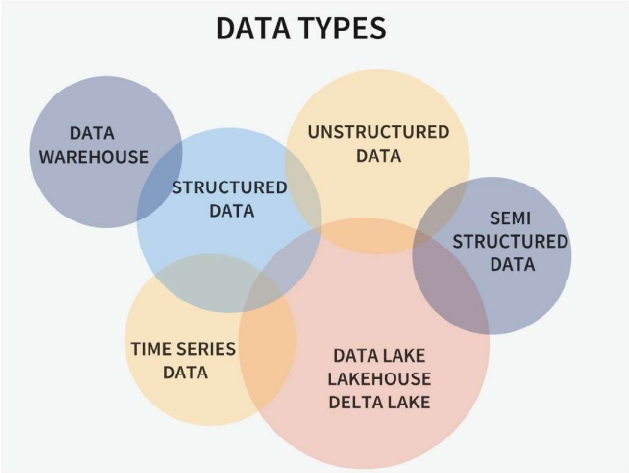


Comparison

Items	Data Lake	Lakehouse	Delta Lake	Data Warehouse
Data Types	Structured, Unstructured, Semi structured	Structured, Unstructured, Semi structured	Structured, Unstructured, Semi structured	Structured
Scalability	Yes: Horizontal Scaling become expensive exponentially	Yes: Horizontal as storage cost are low and compute cost getting cheaper scalability is efficient	Yes: Horizontal as storage cost are low and compute cost getting cheaper scalability is efficient	Possible: Horizonal scaling possible
Performance	High	High	Medium	High
Concurrency	Increase with compute node	Increase with compute node	Increase with compute node	Have limitation with horizontal scaling
Data Access	API, SQL	API, SQL	API, SQL	SQL Only
Governance	Poor governance extra security required	Poor governance extra security required	High row level and column level possible	High row level and column level possible
Security	External Security	External Security	DB level Row and Column	DB level Row and Column
Support	High cost per data stored and processed	High cost per data stored and processed	Low as compared to the Data Lake	High per data processed

Source Data Types Supports and Satisfaction: Tip





Customer satisfaction and coverage increases exponentially with the Lakehouse

CHAPTER 2

Analysis For Implementation Of Data Lake, Lakehouse and Data Warehouse

Data Culture, Organization Data Maturity and Data Road Map

We need to understand the data culture of an organization before we dig further for analysis. Data culture provides information about the organization's data maturity and data-driven capabilities.

"Data culture" refers to a company's ability to use data to make decisions. Companies with a strong data culture have robust analytic capabilities. All the strategies are driven through the analytics information derived from the strong data culture.

Advantages of Data Culture

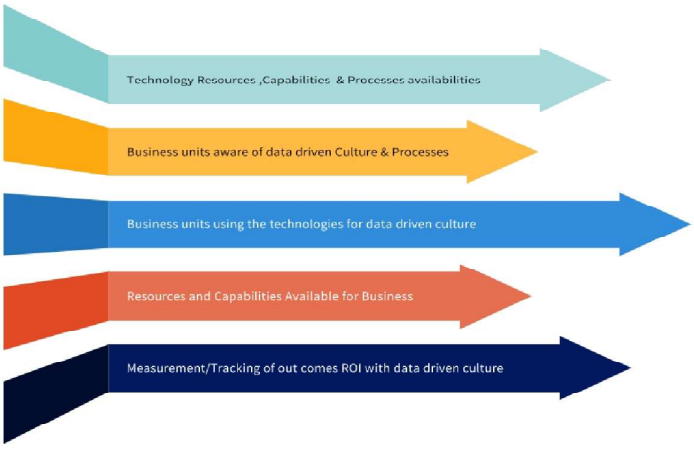
- ✓ Improve the process and understanding of data usage
- ✓ Accelerate customer experience

- ✓ Act as a catalyst for business strategy and decision making
- ✓ Determine the data maturity of an organization
- ✓ Enhance business ROI and creates better business opportunities

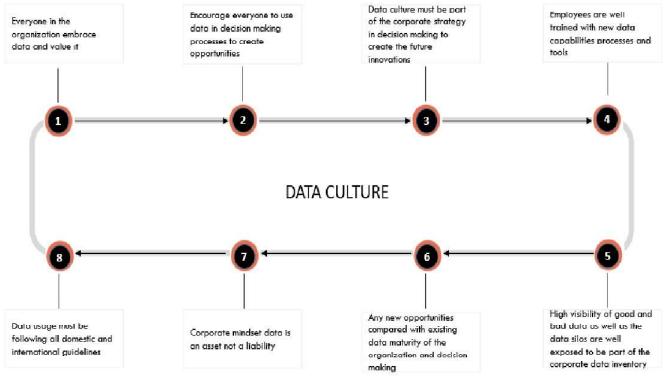
Disadvantages of Data Culture

- ✓ Without proper Data Culture creates trust issues for data
- ✓ Takes longer for decision making and organizational strategy
- ✓ Creates data silos and gaps between the different business units
- ✓ Repetition of duplicate data and data clarity

Data Driven Culture Chart

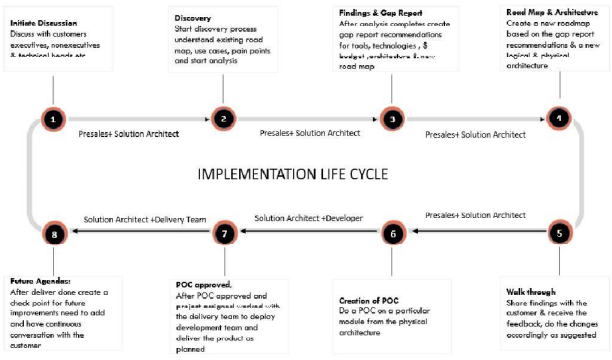


Data Culture Maturity Analysis



Building Blocks Analysis to Deployment:

The below diagram describes the building blocks required to complete a total "Analysis to Delivery" including post-sale supports. Also, each point is described in detail subsequently



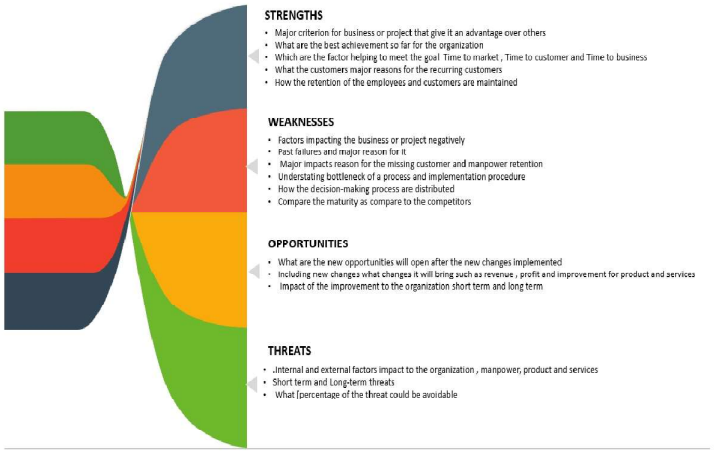
SWOT Analysis

SWOT stands for Strengths, Weaknesses, Opportunities, and Threats. SWOT Analysis is a technique for assessing these four aspects of your organization, business units, or the solution you will be provide.

It is recommended to use the SWOT analysis to assess the current position of your organization before you decide on any new strategies. Here are the points to analyze.

- ✓ Figure it out things positively working for the organization and what are the things impacting negatively.
- ✓ Ask yourself a question about how you will reach the road map keeping client's existing roadmap in mind as well as the road map or solutions you will be recommending to the client.

These could be possible through a SWOT analysis process. So be honest and focus on your people, resources, systems, and procedures.



What Discovery Is

Before we move further, I need to bring to your attention how important it is to understand the current ecosystem, pain points and use cases, etc., eventually to come up with a solution. Recommendations require a thorough analysis and understanding of the current ecosystem which is only possible through a proper analysis process.

This process is usually called "Discovery" I have outlined a summary of the "Discovery" process. This information will give an overall idea of how one should carry forward with the Discovery process. The Discovery is very much crucial to provide a successful recommendation and a solution.

Before someone starts the Discovery process, thoroughly understand the available use cases and what the client is looking for. One example could be that the client wants to see a near-real-time transactional report to determine the

sales by the hour, or they may want to see the available inventory or for the number of customers who order a particular product in the last few hours.

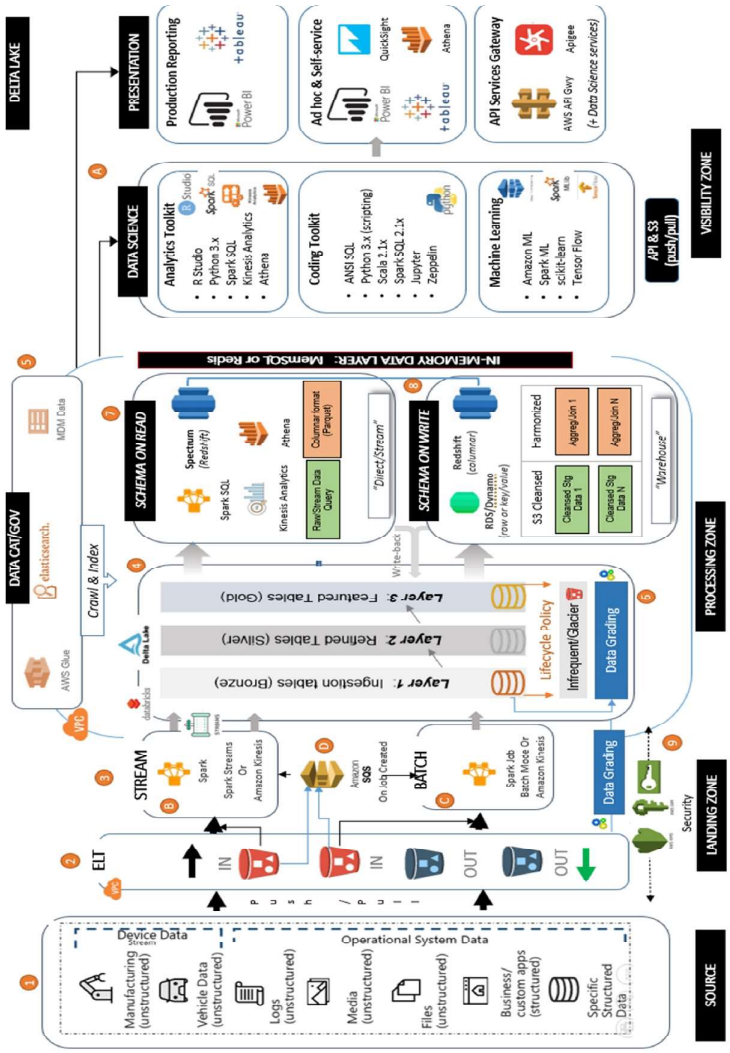
The Discovery process gives the complete outcomes of a project such as the requirements for the tools, technologies, manpower, budget, risks, benefits, road map, logical architecture, physical architecture, a gap report, and implementation plan, etc. I will further go into detail about each step of a discovery process.

Discovery Steps: 

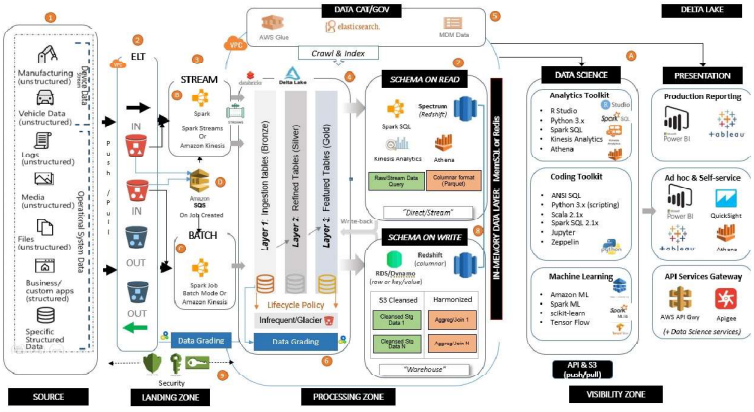
1. Understanding of client's current roadmap and vision
2. Discussion with the business stakeholders (CIO, CFO and CXO), sponsors, executive, mid- level manager, developers, and analysts etc.
3. Collection of use cases need to be implemented and prioritize with development cycle
4. Understanding ecosystems, tools, technologies, manpower, budget, and end customers
5. Gap analysis to determine what the customer have and what the customer wants
6. Determine the timeline, $ figure, tools, technologies, and manpower requirements
7. Creation of a new road map as per the gap report what have and what need to newly introduce
8. Design a Logical architecture

9. Design a Physical (technical) architecture
10. Map use cases to the project road map and technical architecture
11. Arrange a walk-through session with the customer or a workshop
12. After discussion with the customer do the changes suggested during after the walk through or workshop session
13. Prepare a final presentation
14. Divide the total delivery into small modules (tasks and subtasks) and conduct POC
15. After the POC approval by the customer then the project is ready to go for development and delivery cycle

The above task usually takes 6 to 8 weeks based on the project and team size.


DELTA LAKE
PRESENTATION
Production Reporting
Power BI
tableau
Ad hoc & Self-service
Power BI
tableau
QuickSight
Athena
API Services Gateway
AWS API Gwy
Apigee
(+ Data Science services)
DATA SCIENCE
Analytics Toolkit
R Studio
Python 3.x
Spark SQL
Kinesis Analytics
Athena
Coding Toolkit
ANSI SQL
Python 3.x (scripting)
Scala 2.1x
Spark SQL 2.1x
Jupyter
Zeppelin
Machine Learning
Amazon ML
Spark ML
scikit-learn
Tensor Flow
API & S3 (push/pull)
VISIBILITY ZONE
IN-MEMORY DATA LAYER: MemSQL or Redis
DATA CAT/GOV
MDM Data
elasticsearch
AWS Glue
Crawl & Index
SCHEMA ON READ
Spectrum (Redshift)
Spark SQL
Athena
Kinesis Analytics
Columnar format (Parquet)
RawStream Data Query
"Direct/Stream"
SCHEMA ON WRITE
Redshift (columnar)
RDS/Dynamo (row or key/value)
S3 Cleansed
harmonized
Cleansed Sig Data 1
Cleansed Sig Data N
Aggregation 1
Aggregation N
"Warehouse"
Write-back
Layer 3: Featured Tables (Gold)
Layer 2: Refined Tables (Silver)
Layer 1: Ingestion tables (Bronze)
Lifecycle Policy
Infrequent/Glacier
Data Grading
PROCESSING ZONE
databricks
Data Lake
VPC
STREAM
Spark
Spark Streams Or Amazon Kinesis
BATCH
Spark Job Batch Mode Or Amazon Kinesis
Amazon SQS
On Job Created
ELT
IN
OUT
Push
Pull
Data Grading
Security
LANDING ZONE
Device Data
Streams
Manufacturing (unstructured)
Vehicle Data (unstructured)
Operational System Data
Logs (unstructured)
Media (unstructured)
Files (unstructured)
Business/ custom apps (structured)
Specific Structured Data
SOURCE

Physical Diagram Delta Lake with AWS Platform



1. Source Data Batch and Stream
2. Spark Stream Jobs triggered by the Data Brick with micro batch (thirty second) based on this the SQS queued message objects from S3 bucket append to Delta Lake tables by Spark job
3. Spark Batch Jobs triggered by the Data Brick with batch (every predefined hours) based on this the SQS queued new message objects from S3 bucket append to Delta Lake tables by Spark Job
4. For every created object, an event notification is sent to the SQS new log queue
   Based on the predefined schedule of the Spark job data read all new messages mentioned in the new log que
   Spark job append to the Delta table
5. Landing Bucket for IN bucket and OUT Bucket
   after process data ready for consumption
6. Stream data consumption using Spark ,Lambda & Batch mode data using ELT process using ETL tool Talend move to Staging bucket
7. Data get cleansed move from Stage to Cleanse finally to Target DB
8. Data catalog and Master data management using GLUE and Elastic search for search of records
9. Data grading to measure the source data quality and backup policy to archive data
10. Final consumption destination data bases Schema on Read mainly for stream data
11. Final consumption to Data Warehouse or other destination data bases Schema on Write mainly for batch data
12. Security using AIM, KMS etc .
13. Out put visibility layer for Reports & ML

SUMMARY

After going through all the above discussions. it seems the future of data highly depends on the consumption, storing, and processing of the data with different velocities and varieties. An efficient system could be where all types of data are stored in a common storing layer with a standard format, with proper data governance and quality in place. The consumption of data from the common storage area for downstream requirements makes it more convenient. Also, ACID activities on the data are more important for the data analysis. This concludes that a Delta Lake could be have high potential for the data analytics and data visibility; and, meet the upcoming demands for data and analytics. This could provide growth opportunities as well as meet the demands of an organization: Time to Market, Time to Customer, and Time to Business.

Special Thanks to

Mayunthan Nithiyanantham

for the inspiration

A

ACID (Atomicity, Consistency, Isolation, Durability): ACID, 21, 22, 23, 25, 80

API is the acronym for Application Programming Interface: API, 27, 63

E

Extraction Load Transfer: ELT, 16, 41, 48, 49

Extraction Transfer Load: ETL, 6, 14, 16, 17, 19, 25, 26, 41, 48, 49, 51, 52, 68

G

GDPR The General Data Protection Regulation 2016/679 is a regulation in EU law on data protection and privacy in the European Union and the European Economic Area: GDPR-by wiki, 74

M

Microservices: A style of software architecture where complex applications is composed of small, independent services which exchange data and procedural requests; Microservices, 42

S

SQL-Structured Query Language: SQL, 27

Strengths, Weaknesses, Opportunities, and Threats: SWOT, 34

T

Upsert : To insert rows into a database table if they do not already exist, or update them if they do. upsert (plural upserts): Upsert, 24

About the Author

Ajit Dash has spent more than 24+ years in data and analytics in various capacities, led various projects as a Sr. Director Data / Cloud Advisor/ Solution Architect / Cloudy data strategy manager/Advance Analytics/Data Scientist Lead, Pre Sales Lead providing Enterprise and Cross platform integration solutions to various corporations.

His expertise includes strong hands-on analysis and design of Enterprise Solution Architecture, Cloud Advisor, Data Lake, Bigdata, Data Sc, Data warehouse and Data mining, Database Management/Integration, BI Reporting, and Cross-Platform

Domain Expertise: Telecommunication, Biotech, Finance, Banking, Media, Aerospace, Insurance and Technology etc.: (Clients: Fox, Oshkosh, Otis, Travelers, Apple, Qualcomm, IBM, LPL Fin. etc.)

Education:

Ajit Dash holds a Master's degree in General Management from Harvard University

Master's in Computer Information Systems from University of Phoenix

Bachelor's Degree in Electrical Engineering from India

Blog: http: //www.thedataworld.org

US $21.99 CAD $24.99

ISBN 978-1-008-93911-0

9 781008 939110