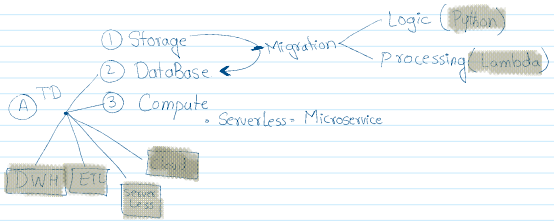


Ⓑ Big Data

Ⓒ New Data

6. ETL
   - DWH - Streaming - DP
   - Tool (Glue, Azure DataBricks)

7. DWH

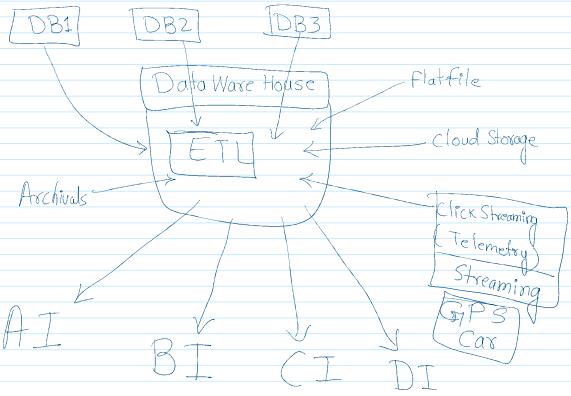
8. Streaming
   Quiz

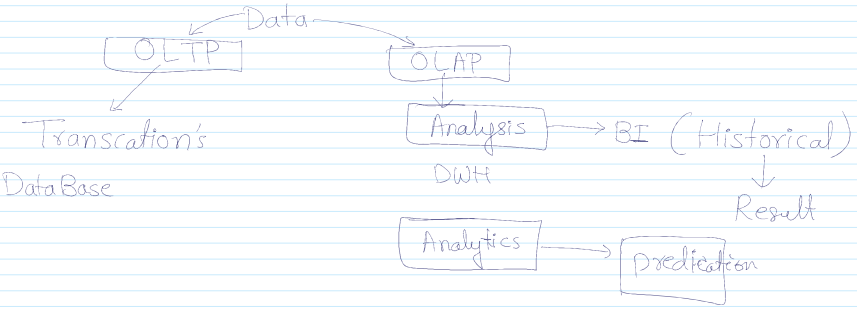
9. Data Pipeline

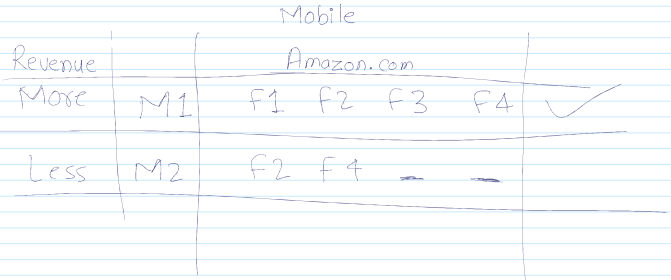
10. API Gateway

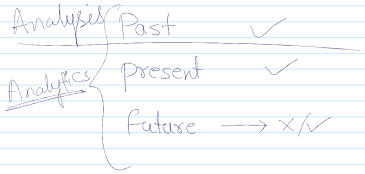
11. Big Data

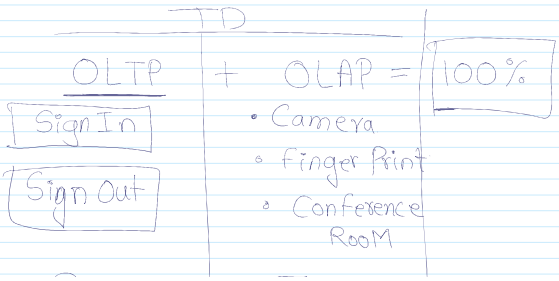
* What is DWH?

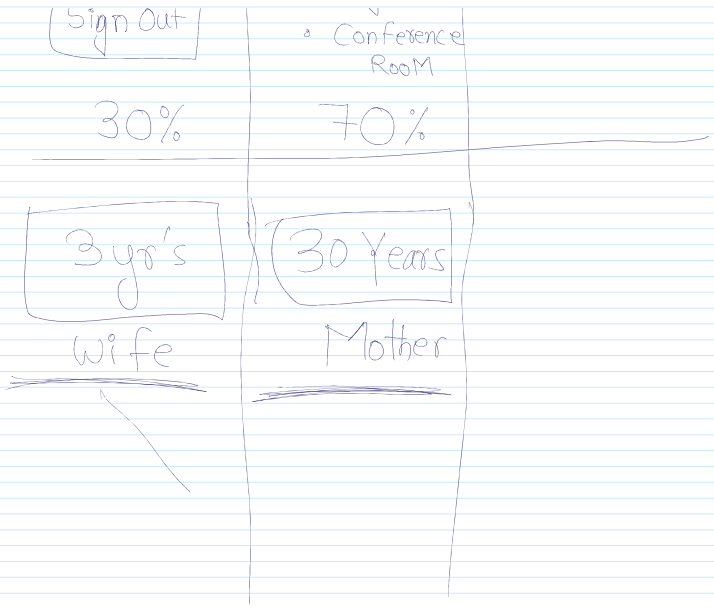


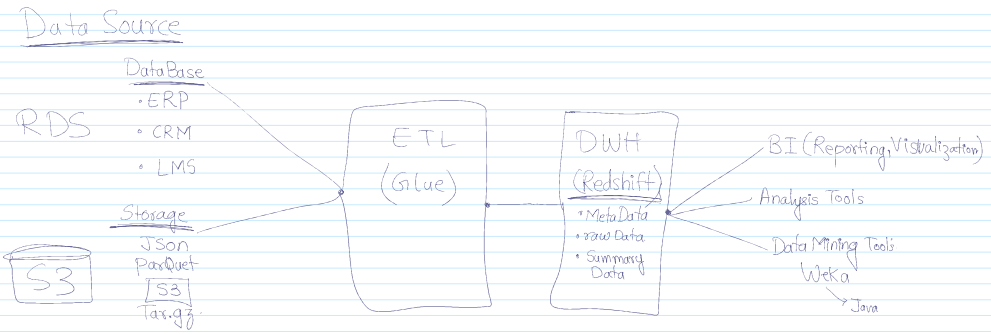


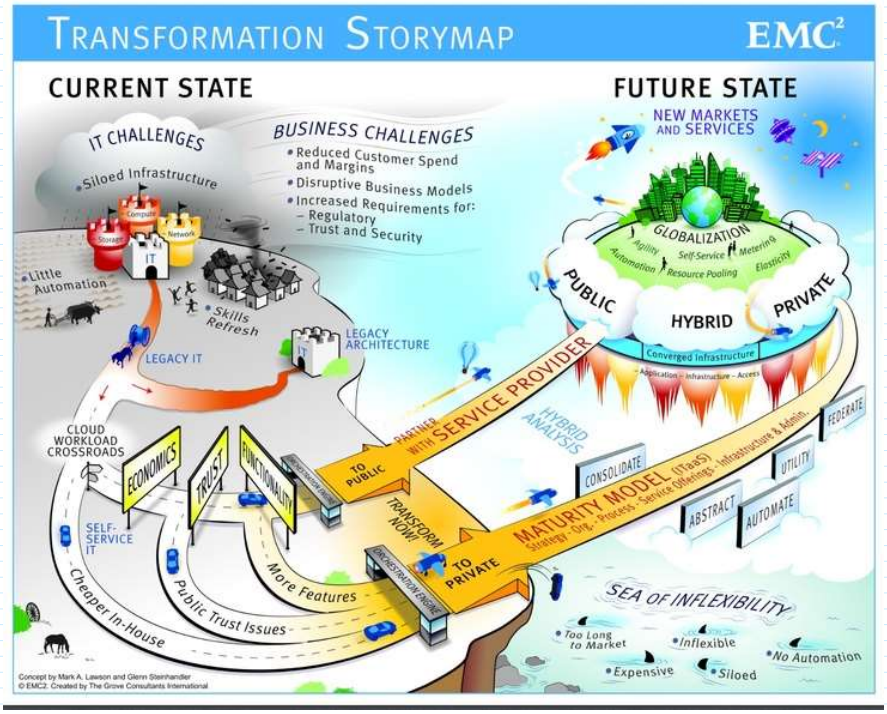


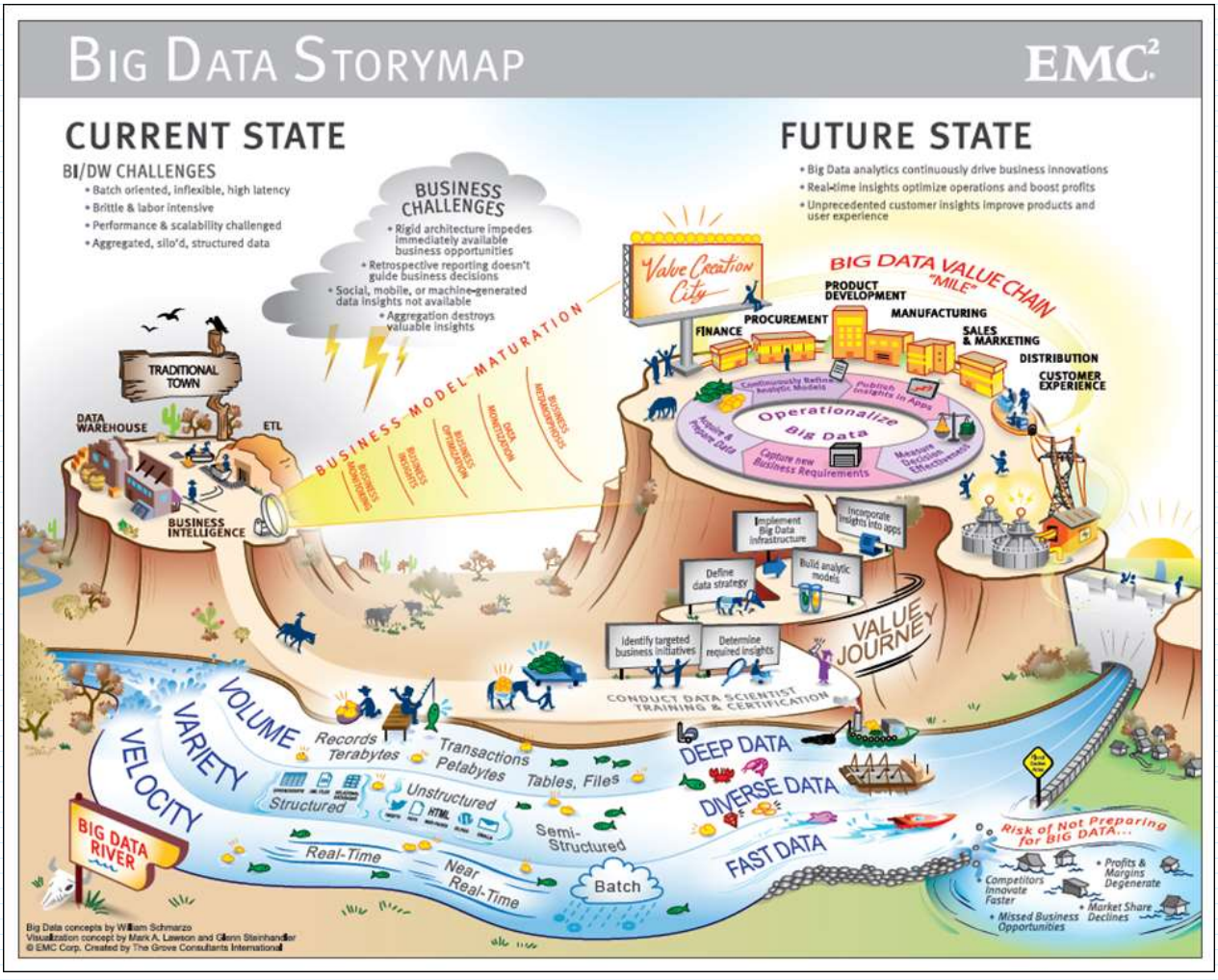


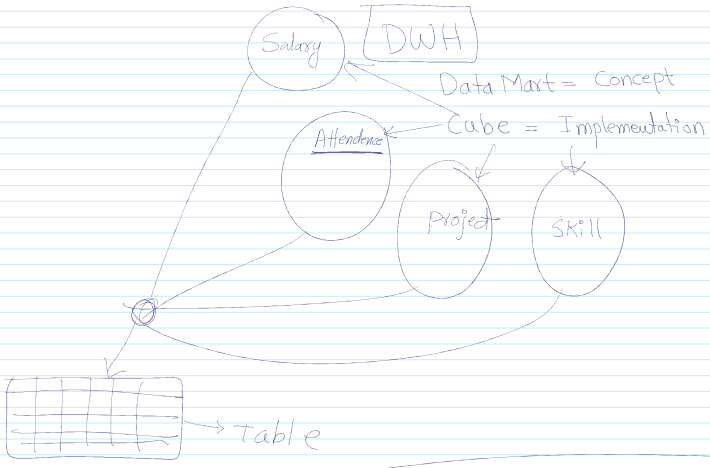


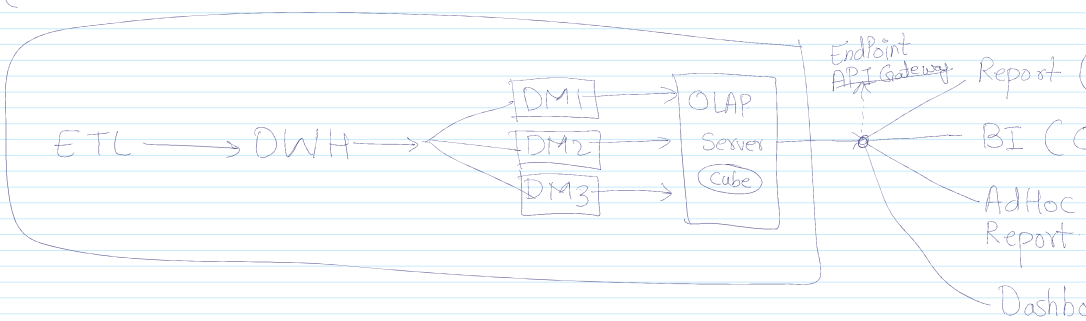

Sign Out
Conference Room
30%
70%
3yr's
30 Years
Wife
Mother


Data Source
DataBase
RDS
• ERP
• CRM
• LMS
Storage
S3
JSon
ParQuet
S3
Tars.gz
ETL
(Glue)
DWH
(Redshift)
• MetaData
• raw Data
• Summary Data
BI (Reporting, Visualization)
Analysis Tools
Data Mining Tools
Weka
Java


TRANSFORMATION STORYMAP
EMC²
CURRENT STATE
FUTURE STATE
IT CHALLENGES
Siloed Infrastructure
Little Automation
Skills Refresh
BUSINESS CHALLENGES
Reduced Customer Spend and Margins
Disruptive Business Models
Increased Requirements for:
– Regulatory
– Trust and Security
LEGACY IT
LEGACY ARCHITECTURE
CLOUD WORKLOAD CROSSROADS
ECONOMICS
TRUST
FUNCTIONALITY
SELF-SERVICE IT
Cheaper In-House
Public Trust Issues
More Features
TO PUBLIC
TO PRIVATE
TRANSFORM NOW!
PARTNER WITH SERVICE PROVIDER
NEW MARKETS AND SERVICES
GLOBALIZATION
PUBLIC
HYBRID
PRIVATE
Converged Infrastructure
HYBRID ANALYSIS
MATURITY MODEL (ITaaS)
CONSOLIDATE
ABSTRACT
AUTOMATE
UTILITY
FEDERATE
SEA OF INFLEXIBILITY
Too Long to Market
Inflexible
No Automation
Expensive
Siloed
Concept by Mark A. Lawson and Glenn Steinhandler
© EMC2. Created by The Grove Consultants International


BIG DATA STORYMAP
EMC²
CURRENT STATE
BI/DW CHALLENGES
Batch oriented, inflexible, high latency
Brittle & labor intensive
Performance & scalability challenged
Aggregated, silo'd, structured data
BUSINESS CHALLENGES
Rigid architecture impedes immediately available business opportunities
Retrospective reporting doesn't guide business decisions
Social, mobile, or machine-generated data insights not available
Aggregation destroys valuable insights
TRADITIONAL TOWN
DATA WAREHOUSE
ETL
BUSINESS INTELLIGENCE
BUSINESS MODEL MATURATION
FUTURE STATE
Big Data analytics continuously drive business innovations
Real-time insights optimize operations and boost profits
Unprecedented customer insights improve products and user experience
Value Creation City
BIG DATA VALUE CHAIN "MILE"
PRODUCT DEVELOPMENT
PROCUREMENT
MANUFACTURING
FINANCE
SALES & MARKETING
DISTRIBUTION
CUSTOMER EXPERIENCE
Operationalize Big Data
Capture new Business Requirements
Measure Decision Effectiveness
Implement Big Data Infrastructure
Incorporate insights into apps
Define data strategy
Build analytic models
Identify targeted business initiatives
Determine required insights
VALUE JOURNEY
CONDUCT DATA SCIENTIST TRAINING & CERTIFICATION
VOLUME
VARIETY
VELOCITY
Records Terabytes
Transactions Petabytes
Tables, Files
Structured
Unstructured
Semi-Structured
Real-Time
Near Real-Time
Batch
BIG DATA RIVER
DEEP DATA
DIVERSE DATA
FAST DATA
Risk of Not Preparing for BIG DATA...
Competitors Innovate Faster
Profits & Margins Degenerate
Market Share Declines
Missed Business Opportunities
Big Data concepts by William Schmarzo
Visualization concept by Mark A. Lawson and Glenn Steinhandler
© EMC Corp. Created by The Grove Consultants International

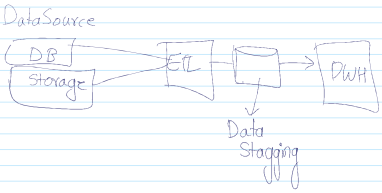




- Data Management System.
- Enable Support
  - BI activities.
  - Analytics
  - Historical Data
  - Report
  - Multiple Data Sources
  - Historical Information
  - Multiple Dimensional
  - DSS

Need DWH

- Business User
- Historical Data
- Strategic Decision
- Repose time.
- Data Consistency



(SSRS)

OBIEE)

oard

	DB	DWH
Processing Type	OLTP	OLAP
Optimized Operations	CRUD transactions	Complex Queries
Data Source	Single	Multiple
Data Model	Normalized	Denormalized
Data time line	Daily to Monthly	Historical
Data Volume	Low to mid	Mid to Large

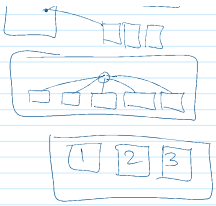
	DB	Streaming
Nature of Data	Storage	Memory
Data Access	Random	Sequentially
Query	one time Queries	Continous Query
	Untimited Storage	Limited
	Low update	High Modify/Insert

AWS Glue

- Discover
- Prepare
- Integrate

} all your data } at Any Scale }
- Analytics
- ML/DL/NLP
- Application Development

Solutions Area	Use Cases	Aws Service's
Adv. Analytics	• Interactive Analytics	• Athena
	• Big Data Processing	• EMR
	• Data Warehouse	• RedShift
	• Real time Analytics	• Kinesis Data Analytics
	• Dashboard & Visualization	• Quick Sight
	• Data Governance.	• Lake Formation • Glue.
	• Object Storage	• S3
	• Block Level	• EBS
	• Serverless Compute	• Lambda
ML	Pred. Analytics & ML	• Sagemaker • DL AMI • Reko • Polly •

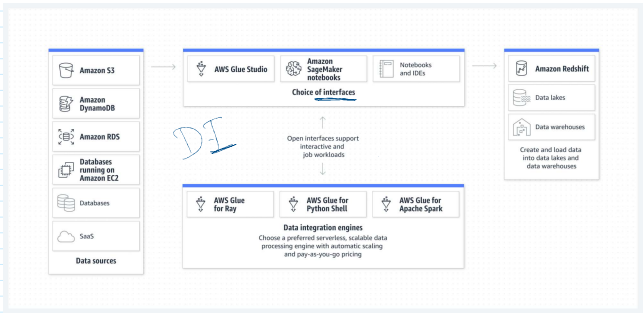


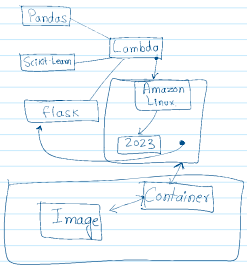
Use Cases of Glue

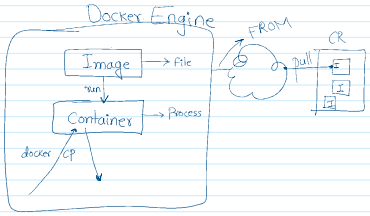
- Event Driven ETL
- Data Integration
- Data Catalog
- No-Code ETL Jobs
- Data Quality
- Data Preparation

Glue is Serverless

Data Integration



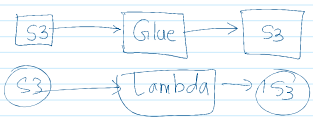




No-Code ETL using Glue

Glue
- Discover & Organize data
- Transform, Prepare & Clean data for analysis
- Build & Monitor Data Pipelines

- Visually transform data
- Build Complex ETL
- Clean & transform Streaming data
- Data Quality
- Built in Job, NB's
- ETL (Edit, Debug, Test)



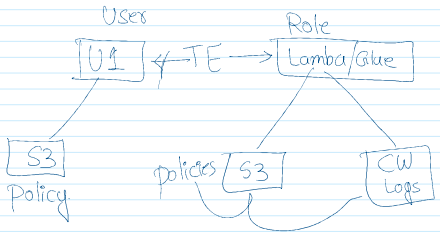
1. I/P
   function
   output

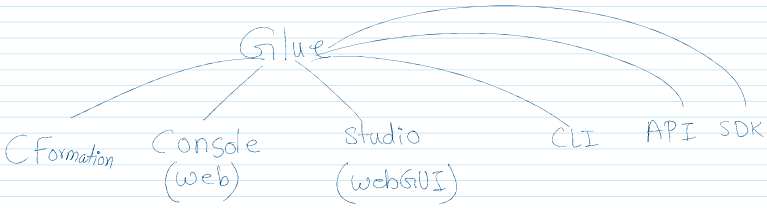
① IAM Role

I/P ② S3 → Bucket → CSV

o/p ③ S3 → Bucket → output

④ Glue → Job





Components

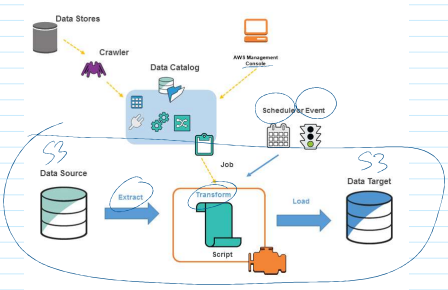
Data Stores

Data Sources

Data Catalog

Job

Data Target

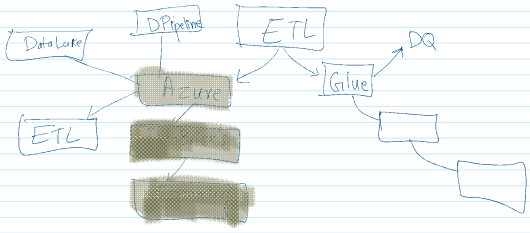


Data Stores:-

- A repository for persistantly storing your data

Data Source

- Trigger/I/p

Job

① Source → Glue Job → Target