Performance Analysis and Speedup of AWS Data Pipeline and Analytics

© JG 2021

Outline

- Motivation
- Why Choose Amazon Web Service (AWS)?
- Serverless Data Analytics Architecture on AWS
- Environment and Security Setup
- Data and Methodology
- Result – Query Performance (MySQL vs Amazon Athena)
- Interactive Dashboard Demo with QuickSight
- Summary

Motivation

❖ **Goal: Performance Analysis and Speedup of AWS Data Pipeline and Analytics**

Data Analytics - Definition

- Extracting insights from data using data analysis and management processes, tools and techniques

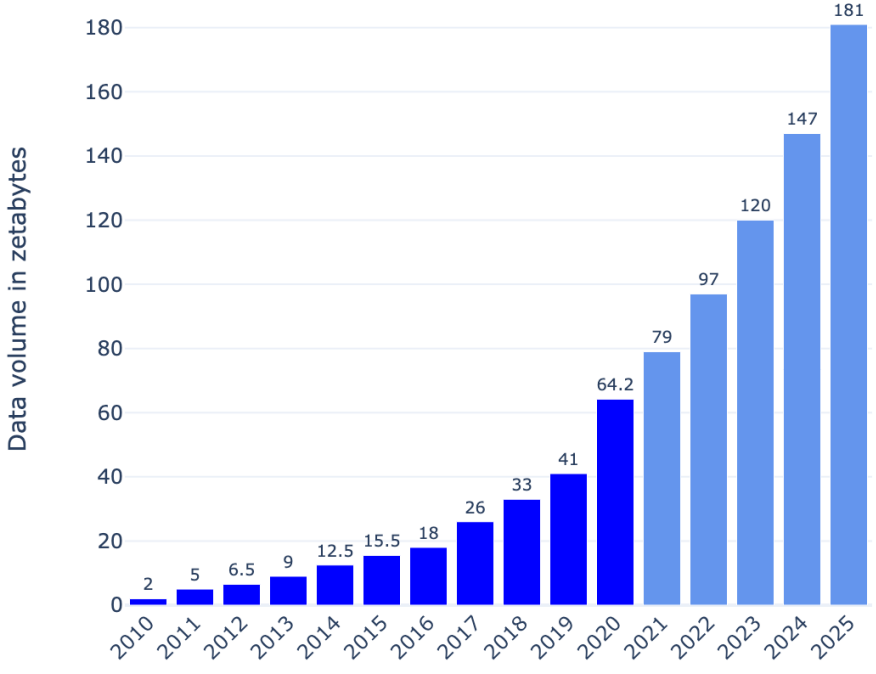
Data Analytics - Importance

- Better and faster decision-making
- To predict future outcomes and enhance business intelligence
- Better operational efficiency
- To improve customer service

Data Analytics - Challenges

- Exponential growth of data with time
- Inefficient data management due to large size, complexity, variety, and variability
- Data velocity − speed of data generation and process to meet demands
- Open-source data analytics tools to meet enterprise security standard
- Processing power − latency and cost of data processing

Worldwide data volume from 2010 to 2025, increasing rapidly − reaching 64.2 zettabytes in 2020 and projected to grow more than 180 zettabytes by 2025.



With the rapid growth of the data volume, storage capacity is projected to increase, growing at a compound annual growth rate of **19.2 %** over the forecast period from 2020 to 2025.

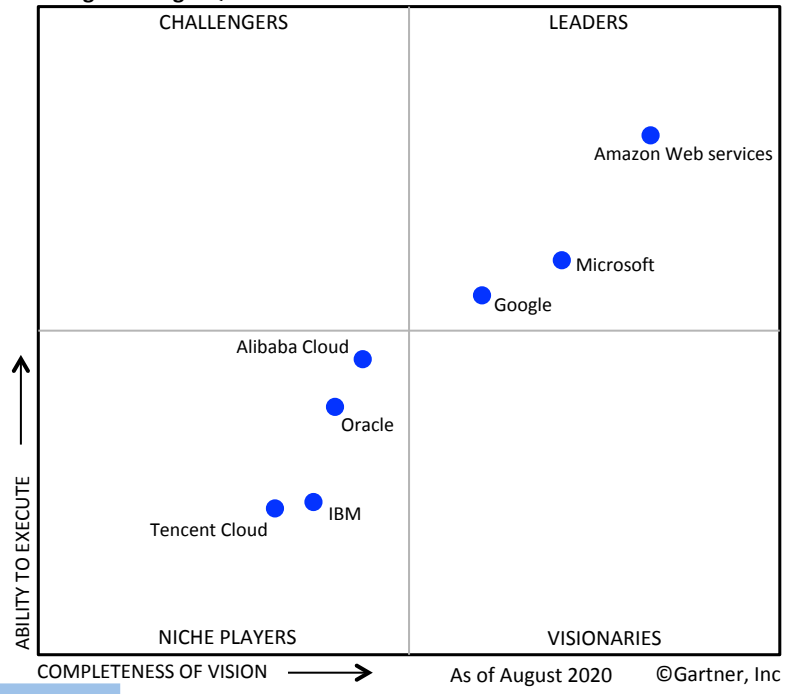
[Statista 2021]

Why Choose AWS for Data Analytics?

❖ Cloud Infrastructure and Platform Services Reviews and Ratings: Ratings (# of reviews) [Gartner peerinsights]

	AWS	Google	Microsoft	IBM
User Management	4.4 (2593)	4.4 (750)	4.3 (1516)	4.2 (204)
Security and Compliance	4.6 (2647)	4.6 (777)	4.4 (1542)	4.3 (211)
Architecture Flexibility	4.5 (238)	4.5 (136)	4.3 (157)	4.1 (57)
Big Data Enabled	4.4 (**293**)	**4.5** (160)	4.3 (180)	4.2 (66)
Developer Services	4.4 (2629)	4.4 (774)	4.4 (1536)	4.1 (210)
Enterprise Integration	4.4 (2571)	4.3 (762)	4.3 (1492)	4.1 (201)
Resilience	4.5 (293)	4.4 (161)	4.3 (180)	4.0 (66)
Ease of Deployment	4.4 (**2431**)	**4.5** (699)	4.3 (1501)	4.1 (229)
Ease of Integration using Standard APIs and Tools	4.4 (**2434**)	**4.5** (696)	4.3 (1504)	4.1 (228)
Pricing Flexibility	4.1 (**2233**)	**4.4** (588)	4.0 (1369)	3.9 (218)
Overall Peer Rating	4.6 (2722)	4.5 (816)	4.4 (1633)	4.3 (268)

Figure: Magic Quadrant for cloud Infrastructure and Platform Services



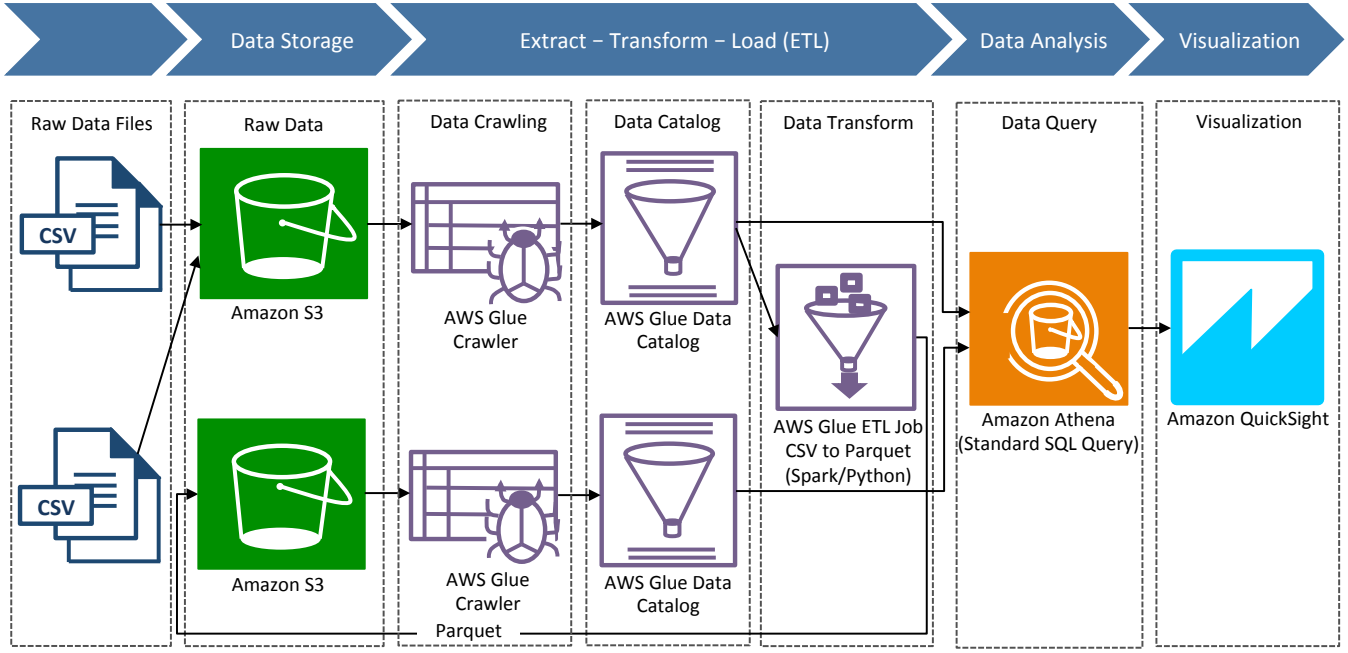
	AWS	Google	Microsoft
# of Services	200+	60+	100+
Availability Zones	66 zones + 12 coming	20 regions + 3 coming	54 regions, 140 country
Market Shares,	32%	7%	19%

[Ref: Intellipaat]

According to Gartner report, AWS is leading with the highest score in both axes of measurement – Ability to Execute and Completeness of Vision.

[Gartnet Report 2020]

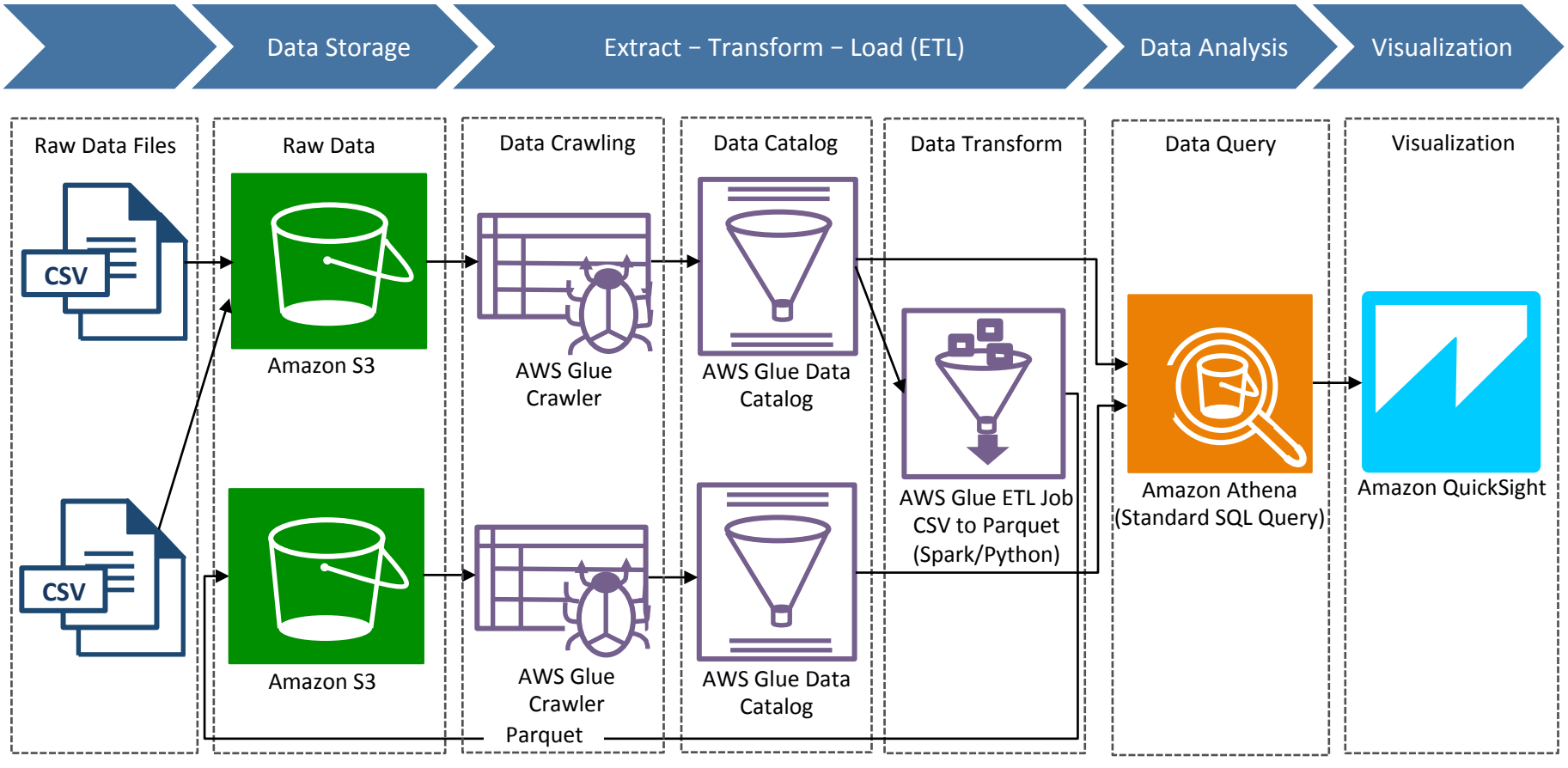
Serverless Data Analytics Architecture on AWS



- Store and analyze large volume of data with high durability, security, and scalability
- In serverless environment, no infrastructure needed to manage and monitor the performance
- Data size reduced by transforming raw CSV data into Parquet (around 5 to 1 compression), resulting in reduced query cost

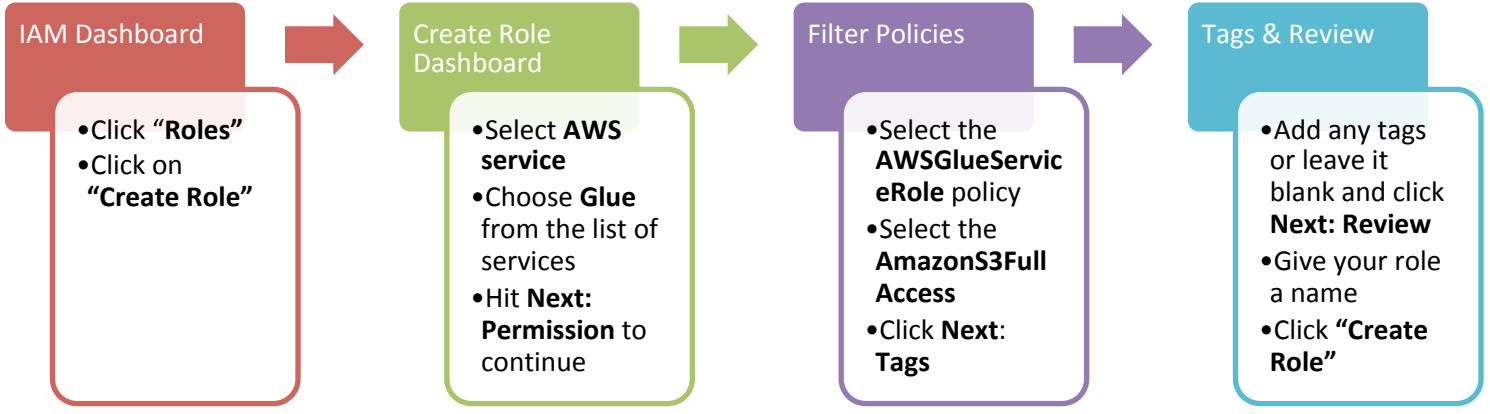
Amazon S3	Amazon Glue	Amazon Athena	Amazon QuickSight
• Unlimited object storage • Structured, semi-structured • Cheap, reliable • Cost effective ($0.023 per GB) • 99.999999999% durability • Standard S3 storage - 99.99% availability	• Completely managed serverless ETL service • Create a unified catalog to find data across multiple data stores • Simple, scalable and faster data integration • Cost effective – pay-as-you-go pricing	• A serverless interactive query service • Query data in Amazon S3 using standard SQL • Faster - executes queries in parallel • Easy to quickly analyze large-scale dataset • Cost effective – only pay for querying data	• A scalable, serverless, embedded, machine-learning (ML) powered BI service • Fully integrated AWS and other third-party data sources • Faster navigation with SPICE • Cost effective – pay-as-you-go pricing

Serverless Data Analytics Architecture on AWS



Environment and Security Setup

- Create and activate an AWS account [see instructions]
- Create an **Amazon Identity and Account Management** (IAM) Role to give
  - S3 bucket access permission
  - Glue service permission for crawling and transforming data

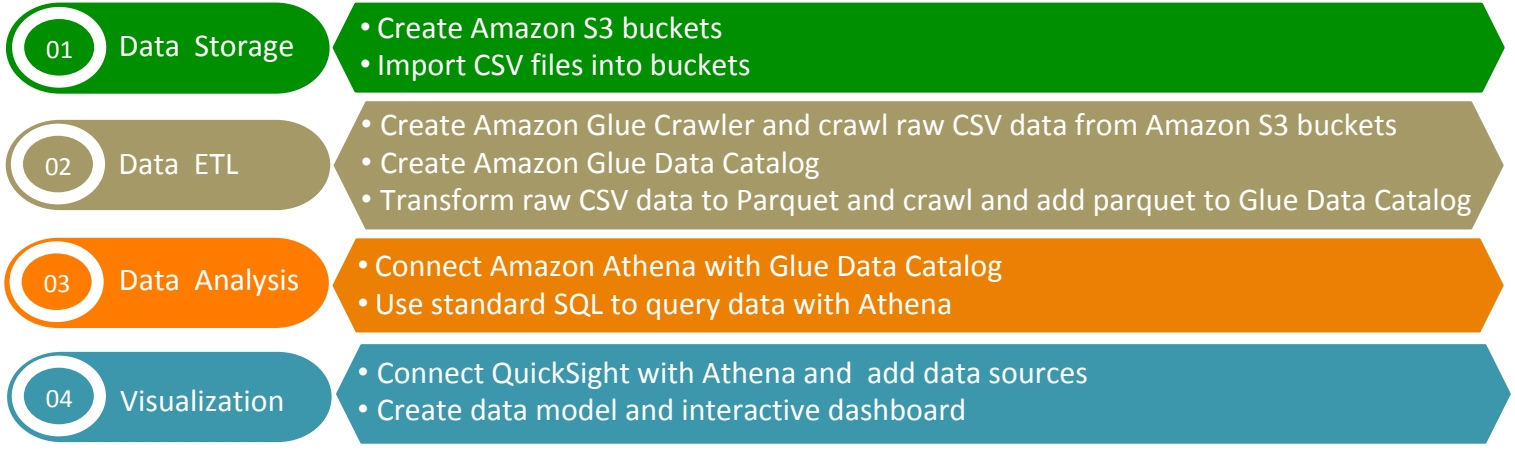


Data and Methodology

❖ **Data Overview:**

Data	Size	Description	Source
City Temperature	140MB	Daily temperature of major cities of the world	kaggle
Electronics Store	360MB	Purchase data from April 2020 to November 2020 from a large home appliances and electronics online store	kaggle
Customer complaints	891MB	Real world complaints data received about financial products and services	kaggle

❖ **Methodology:**



Query Performance (Measure: Data Scanned and Runtime)

Query-A: SELECT Distinct column FROM table;
Query-B: SELECT Count(*) FROM table where column='value';

Table3: Query performance (data scanned): Raw vs Parquet using AWS Athena

Data	Size	SQL Query	Raw Data	Parquet Data	Decrease
City Temperature	140MB	Query-A	140MB	3.18KB	99.99%
		Query-B	140MB	5.26KB	99.99%
Electronics Store	360MB	Query-A	360MB	2.64MB	99.26%
		Query-B	360MB	6.71MB	98.14%
Customer complaints	891MB	Query-A	891MB	894.48KB	99.90%
		Query-B	891MB	2.31MB	99.74%

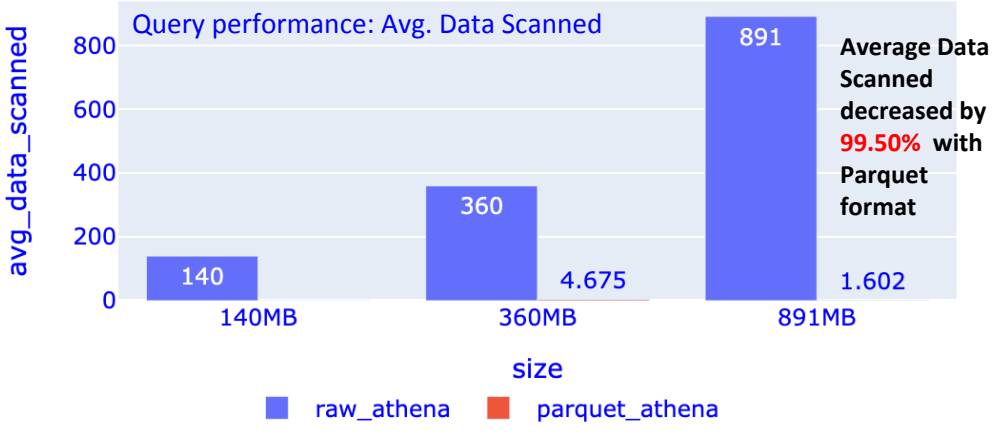
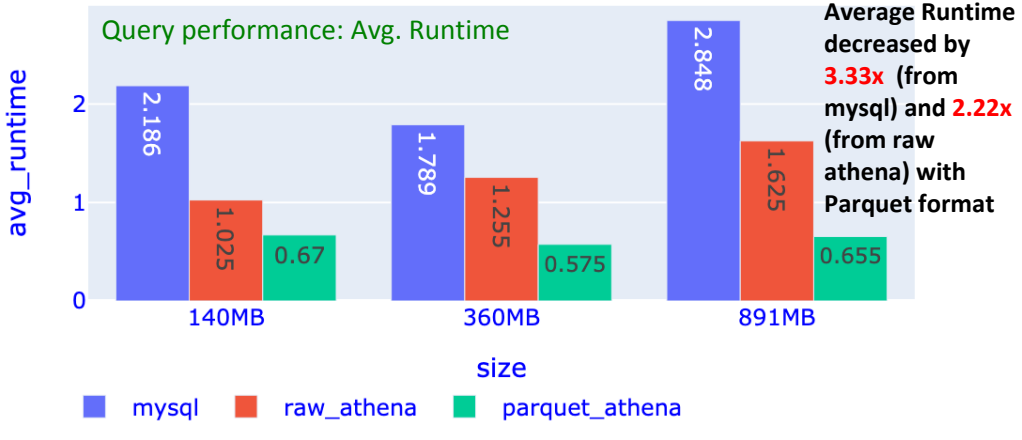


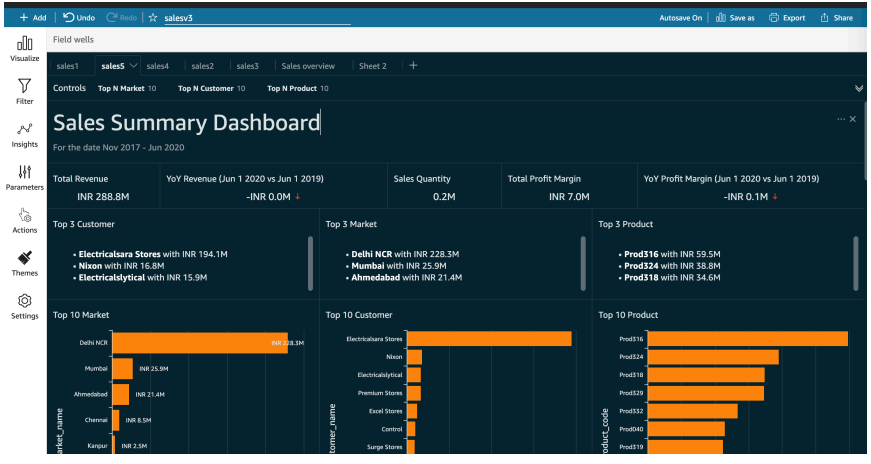
Table4: Query performance (runtime): Raw vs Parquet data with AWS Athena

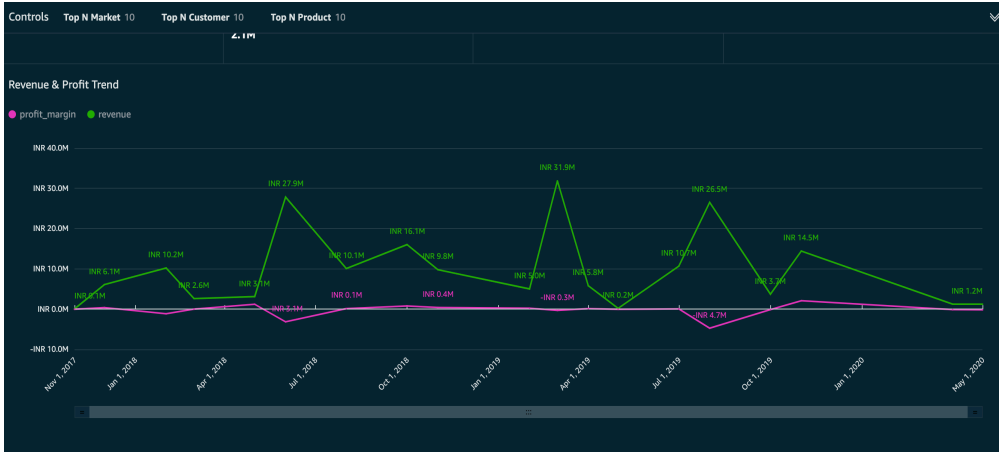
Data	Size	SQL Query	MySQL	Raw data (Athena)	Parquet (Athena)
City Temperature	140MB	Query-A	2.526s	1.14s	0.74s
		Query-B	1.845s	0.91s	0.6s
Electronics Store	360MB	Query-A	1.759s	1.16s	0.65s
		Query-B	1.819s	1.35s	0.5s
Customer complaints	891MB	Query-A	2.490s	2.29s	0.71s
		Query-B	3.205s	0.96s	0.6s

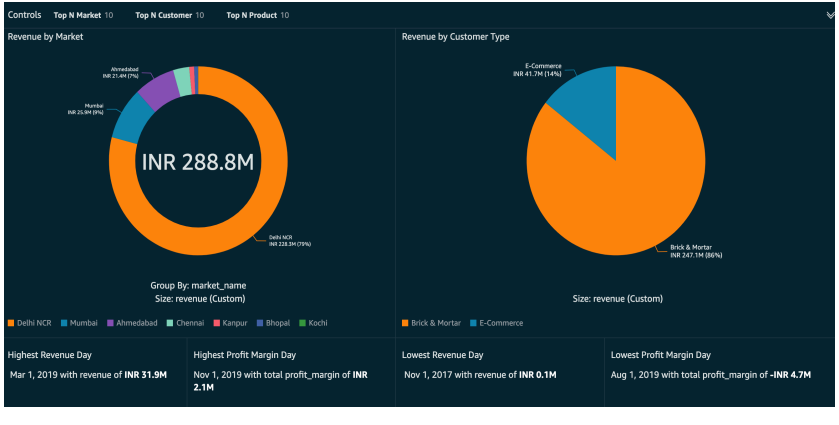


AWS QuickSight Dashboard Demo

❖ Sales Data insights dashboard using AWS QuickSight [data source: github/codebasics].







Details

market_name	zone	customer_name	customer_type	product_code	quantity	cost_price	revenue	profit_margin
Chennai	South	Surge Stores	Brick & Mortar	Prod040	3,400	INR 2.35M	INR 2.3M	-INR 0.1M
Delhi NCR	North	Electricalsara Stores	Brick & Mortar	Prod040	7,851	INR 13.65M	INR 14.4M	INR 0.7M
Delhi NCR	North	Info Stores	Brick & Mortar	Prod040	152	INR 0.24M	INR 0.2M	INR 0.0M
Delhi NCR	North	Nixon	E-Commerce	Prod040	1,006	INR 1.63M	INR 2.0M	INR 0.3M
Delhi NCR	North	Premium Stores	Brick & Mortar	Prod040	2,336	INR 2.89M	INR 3.0M	INR 0.1M
Kanpur	North	Acclaimed Stores	Brick & Mortar	Prod040	563	INR 1.07M	INR 1.0M	-INR 0.1M
Chennai	South	Surge Stores	Brick & Mortar	Prod159	3,418	INR 1.42M	INR 1.4M	-INR 0.1M
Delhi NCR	North	Electricalsara Stores	Brick & Mortar	Prod159	10,318	INR 11.28M	INR 11.6M	INR 0.3M
Delhi NCR	North	Info Stores	Brick & Mortar	Prod159	198	INR 0.19M	INR 0.2M	INR 0.0M
Delhi NCR	North	Nixon	E-Commerce	Prod159	1,404	INR 1.82M	INR 1.9M	INR 0.1M
Delhi NCR	North	Premium Stores	Brick & Mortar	Prod159	1,954	INR 1.65M	INR 1.6M	INR 0.0M
Kanpur	North	Acclaimed Stores	Brick & Mortar	Prod159	445	INR 0.57M	INR 0.5M	-INR 0.1M
Chennai	South	Surge Stores	Brick & Mortar	Prod304	477	INR 0.19M	INR 0.2M	-INR 0.0M
Delhi NCR	North	Electricalsara Stores	Brick & Mortar	Prod304	17,165	INR 13.69M	INR 14.6M	INR 0.9M
Delhi NCR	North	Info Stores	Brick & Mortar	Prod304	65	INR 0.02M	INR 0.0M	INR 0.0M
Delhi NCR	North	Nixon	E-Commerce	Prod304	1,161	INR 1.37M	INR 1.2M	-INR 0.1M
Delhi NCR	North	Premium Stores	Brick & Mortar	Prod304	2,855	INR 1.81M	INR 1.8M	INR 0.0M
Kanpur	North	Acclaimed Stores	Brick & Mortar	Prod304	1	INR 0.00M	INR 0.0M	-INR 0.0M
Total					**207,470**	**INR 281.80M**	**INR 288.8M**	**INR 7.0M**

Summary of Result and Achievement

Result:

- Implemented serverless data pipeline on AWS for performance analysis and speedup
- Performed data analysis and visualization
- Stored CSV data (**140**MB, **360**MB, **891**MB) in Amazon S3 buckets
- Transformed raw data into parquet format and created Glue Data catalog
- Performed SQL queries on Cataloged data (both raw and parquet) using Amazon Athena
- Analyzed and compared query performance: MySQL vs AWS Athena
- Created Dashboard using Amazon QuickSight

Achievement:

1. Data size reduced by **80.46%** by transforming raw CSV data into Parquet (around 5 to 1 compression), resulting in reduced query cost
2. Average Runtime reduced by **3.33x** (from MySQL) and **2.22x** (from raw athena) with parquet data
3. Average Data Scanned reduced by **99.5%** with parquet data