Examples of Real World Analytics Solutions and Data Teams in Western Companies

by Dmitry Anoshin, Data Engineer

SmartData

2023

Outline

- About myself
- What is Analytics and Data Engineering
- Real Analytics Solutions Examples - Teams, Tools and Use Cases
- Career path vs Compensation vs Cost of Living (my opinion)

University of Victoria

Teaching Fundamentals of Cloud Computing, Analytics and AI since 2021

Rock Your Data

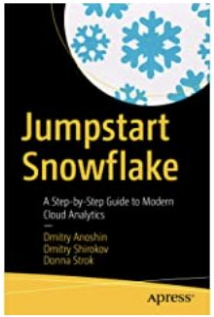
Cloud Analytics Consulting since 2017

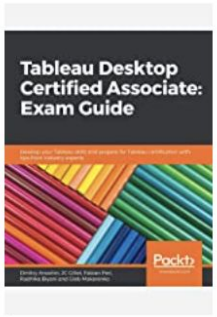
Free Data Engineering and BI data school in RUS with 8000+ students

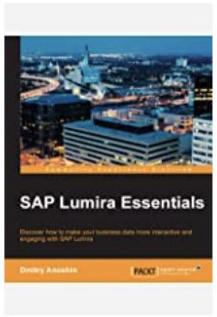
Surfalytics

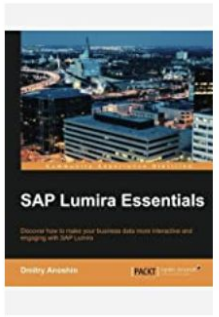
Data Engineering and Analytics IC coaching

- 13 + years in Data Engineering & Analytics as IC
- Moscow, Montenegro, Seattle, Boston, Vancouver
- 5 years @Amazon, 3 years @Microsoft Azure Data & AI
- Tableau, Snowflake, Databricks and etc. user groups and meetups
- Speaker on biggest data conferences
- Author of 7 analytics book with avg rating 4.5



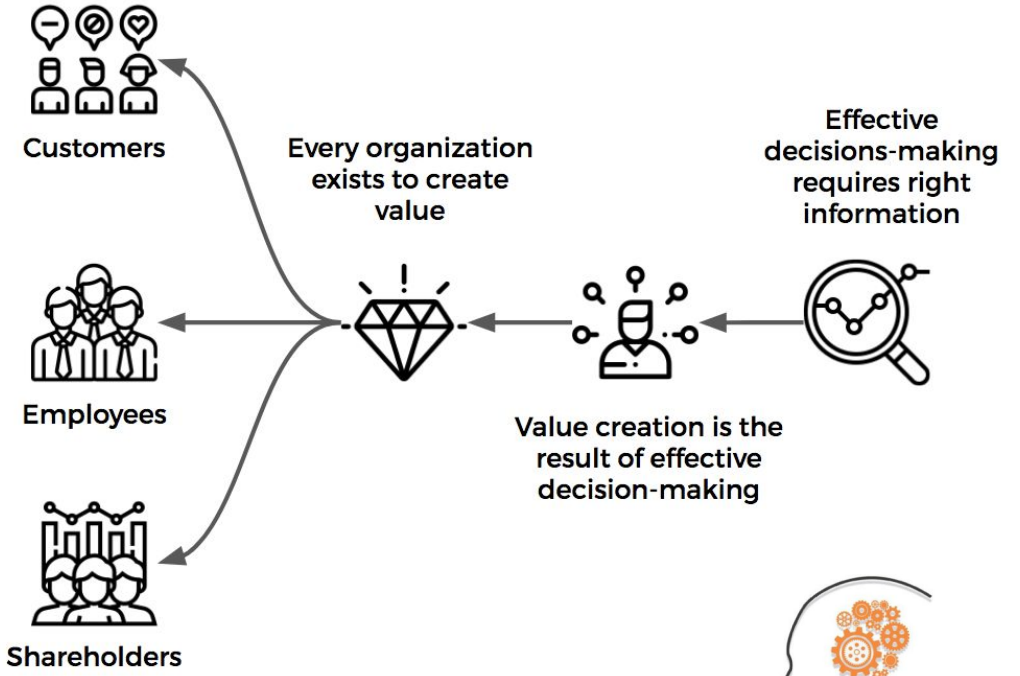






What is Analytics?

- Increase Revenue
- Decrease Cost
- Mitigate Risks
- Research new markets and products
- Validate Hypothesis



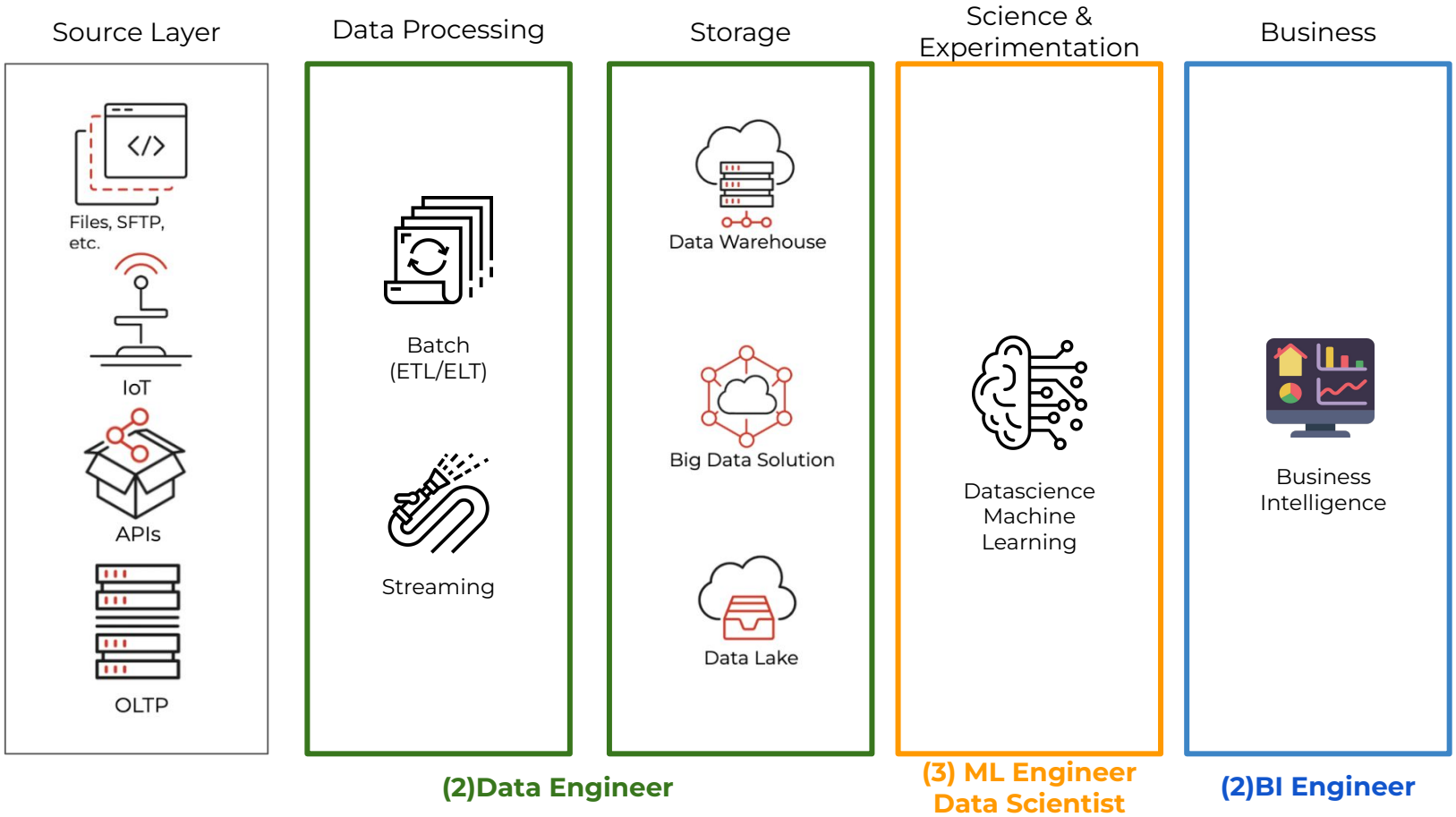
HYPER

Changing the way you think about, plan, and execute
Business Intelligence
for real results, real fast!

Gregory P. Steffine

Key Layers and roles in Data Stack

(1) Manager(s).



What is Data Engineering?

IBM: Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret. Their ultimate goal is to make data accessible so that organizations can use it to evaluate and optimize their performance.

Real Python: The ultimate goal of data engineering is to provide organized, consistent data flow to enable data-driven work

CIO: Data engineers are responsible for finding trends in data sets and developing algorithms to help make raw data more useful to the enterprise.

Dremio: Data engineering helps make data more useful and accessible for consumers of data. To do so, ata engineering must source, transform and analyze data from each system.

Gartner: Data engineers play a key role in building and managing data pipelines, and promoting data and analytics use cases to production (in line with business processes).

Microsoft: Data Engineers help stakeholders understand the data through exploration, and they build and maintain secure and compliant data processing pipelines by using different tools and techniques.

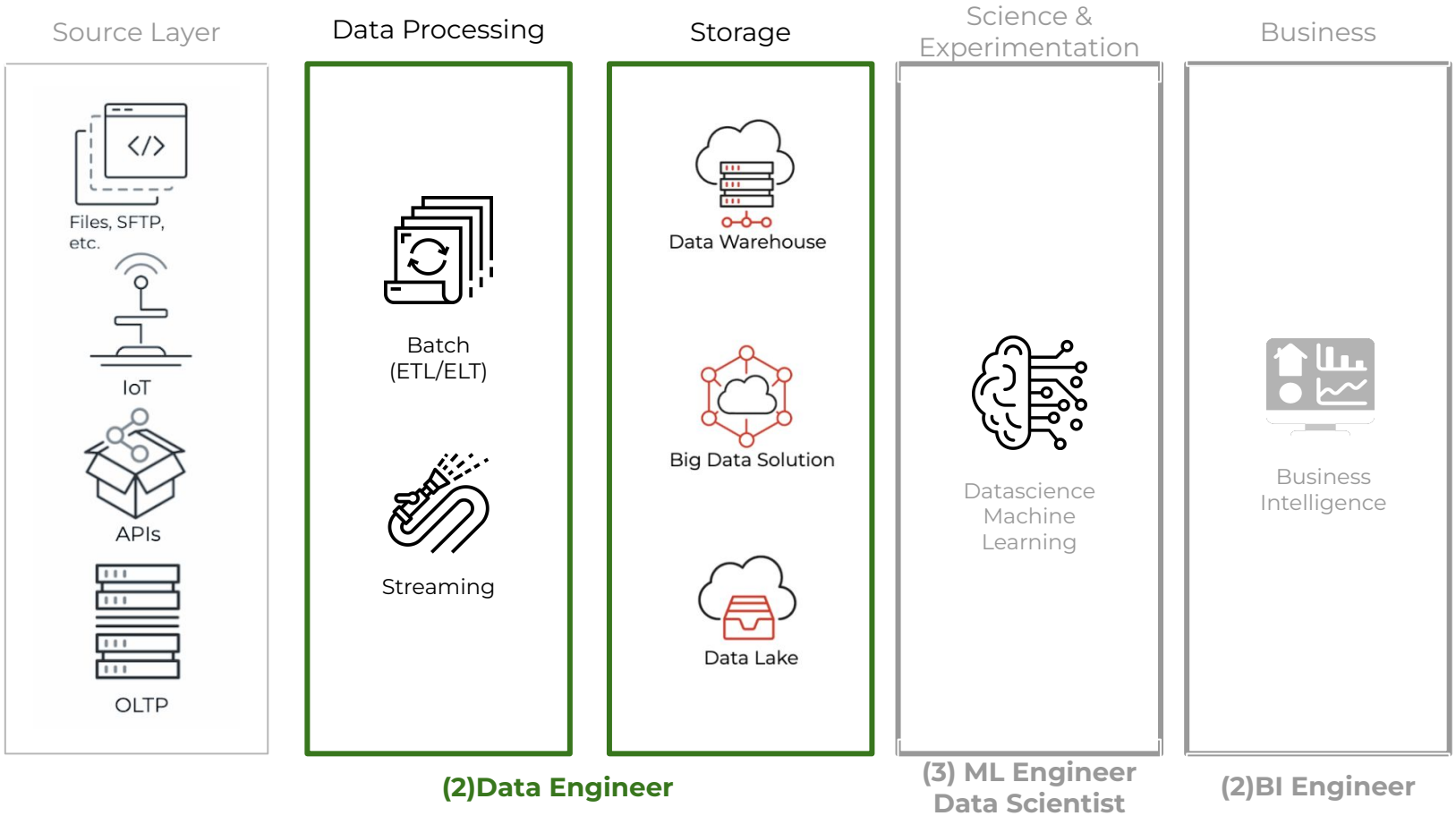
Amazon: Data Engineers tackle some of the most complex challenges in large-scale computing. Most of the work they do involves storing and providing access to data in efficient ways.

TL;DR:

Data engineering makes data useful and accessible for consumers by building secure and scalable data infrastructure.

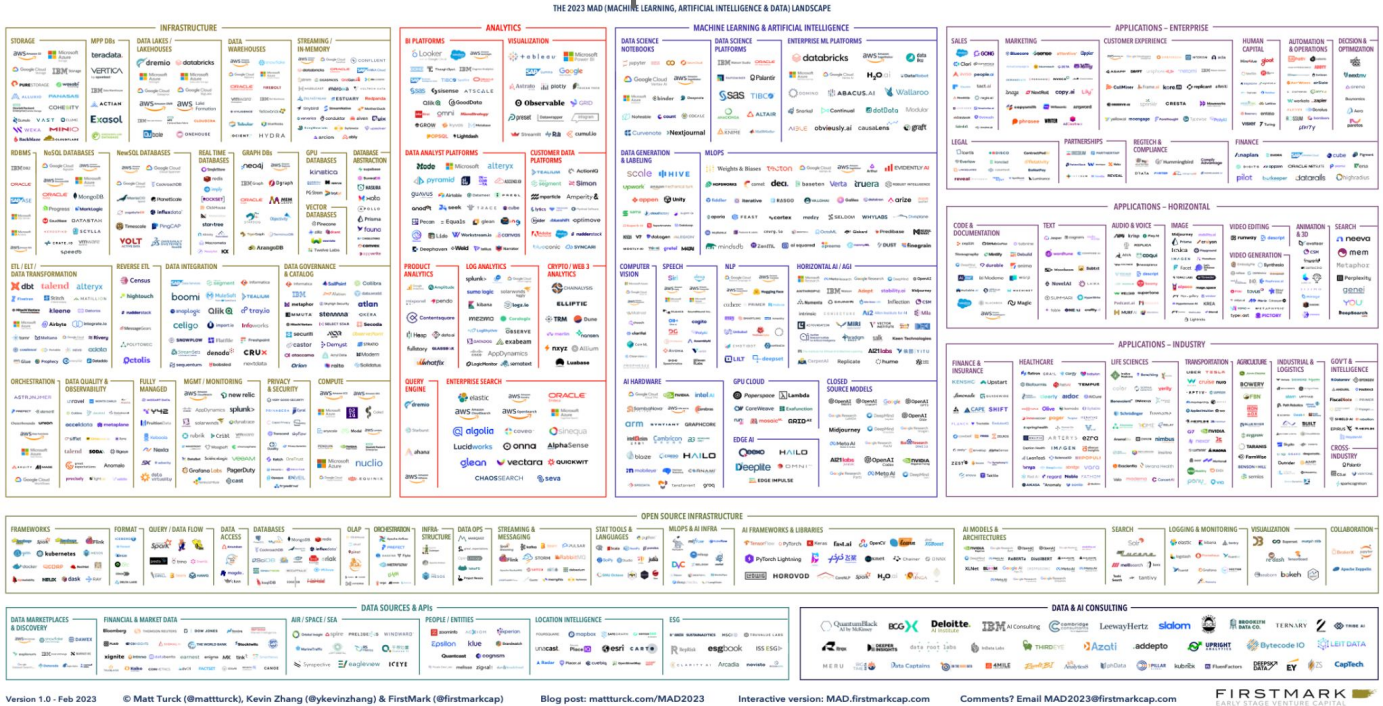
DE Key Layers

(1) Managers.



DE Challenges

1. Data and AI Landscape 2023



2. Same could exist for Trainings and Courses

What will we review for each project?

- Team setup
- Tools for ETL, DW, BI, ML/DS
- Challenges

Case 1: Migration of Oracle Data Warehouse to AWS Redshift

Company:

- Amazon Subsidiary Abebooks - marketplace of used books and collectables
- 400 employees

Data Team:

- 2 Oracle DBAs (team supported back-end OLTP)
- Software Engineering Manager (SDM)
- 1 Data Engineer (who is it in 2016?!)

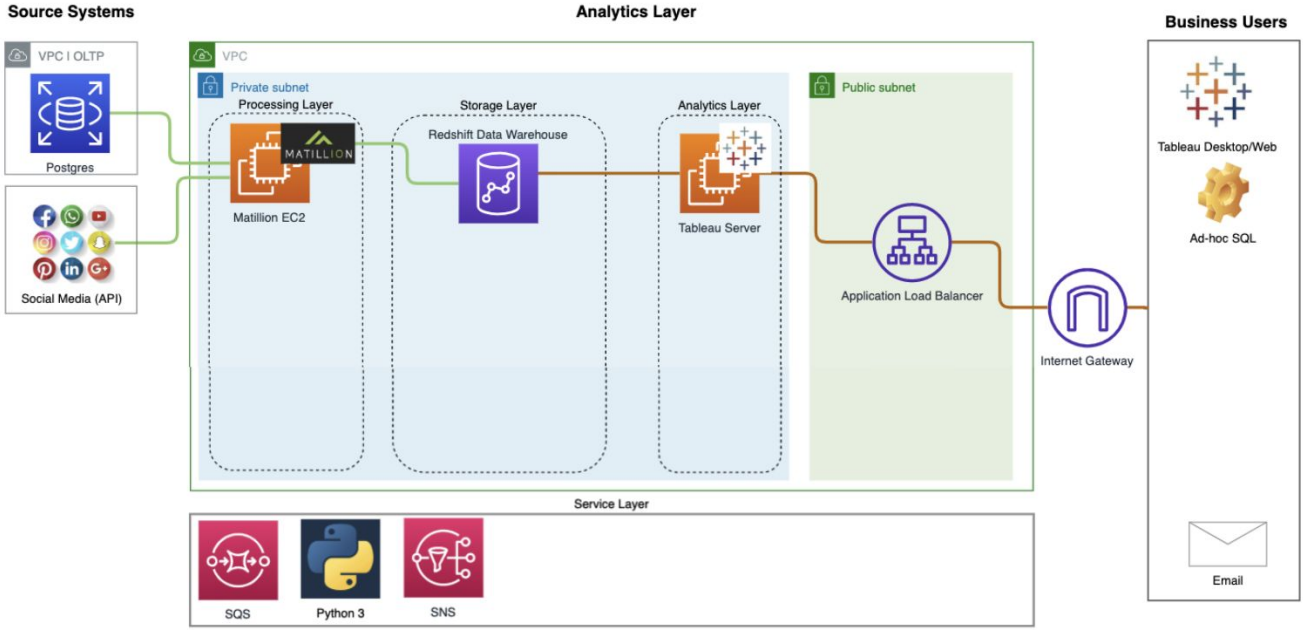
Starting State:

- Oracle DW
- PL/SQL + cron
- Excel + Crystal Reports (scheduler)

Use cases:

- Sales and Financial reporting
- Financial reconciliation
- Marketing Analytics (attribution model, channel performance, user segmentation)
- Product/Category Analytics

Amazon Redshift, Matillion ETL, Tableau



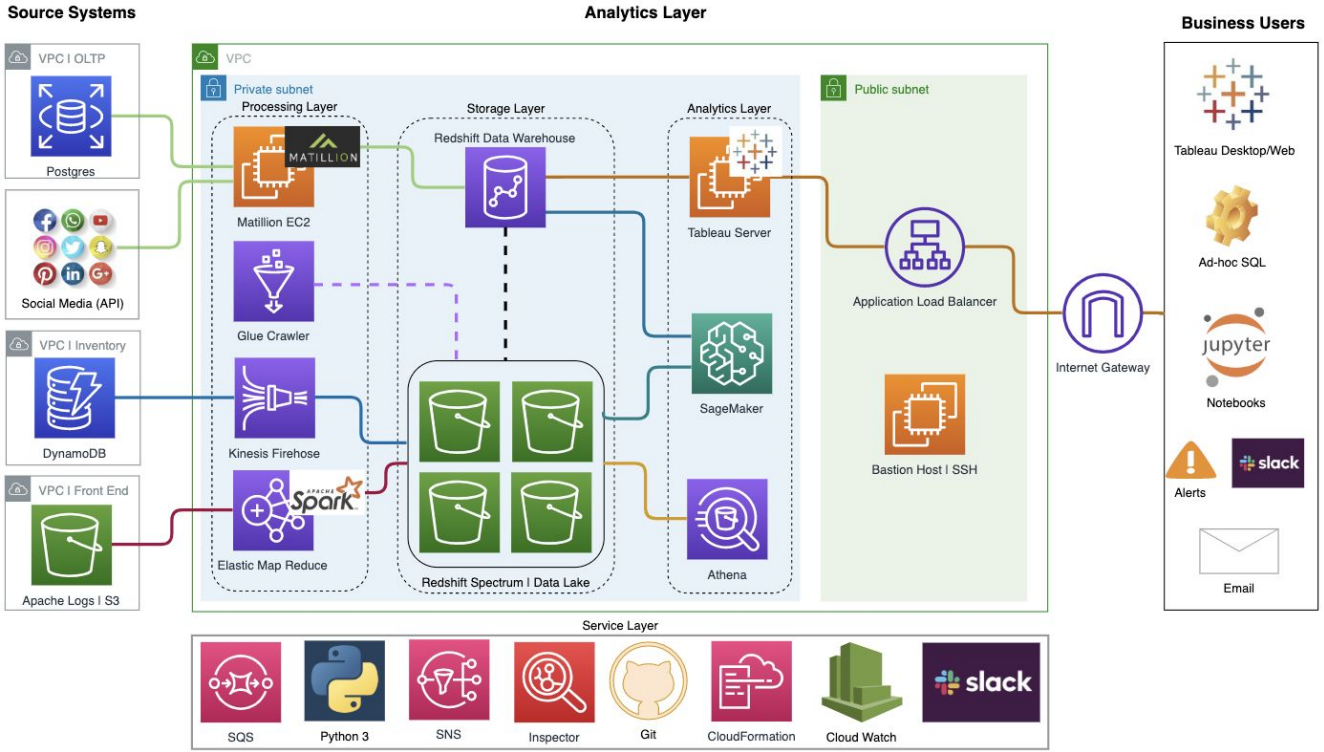
Timeline:

- with 3 people team it took ~8 months to move DW and ETL pipelines to AWS.
- Another 4 months for moving reports to Tableau Server

Solution:

- Redshift DW
- Tableau
- Matillion ETL

Adding “Big Data” component for Web logs and DynamoDB streams for inventory database



Solution:

- AWS Elastic Map Reduce (EMR) + Spark solved performance bottleneck for Redshift ETL
- Backend database in DynamoDB, allowing stream CDC with DynamoDB streams and deliver to S3 via Kinesis

Case 2: Product Analytics Solution on AWS

Company:

- Amazon Alexa, Natural Language Understanding team, internal data products

Team:

- 1 Data Engineer
- 2 BI Engineer
- 1 Front End Engineer
- 2 SDE
- 2 Technical Product Managers
- 1 Sr. Product Manager

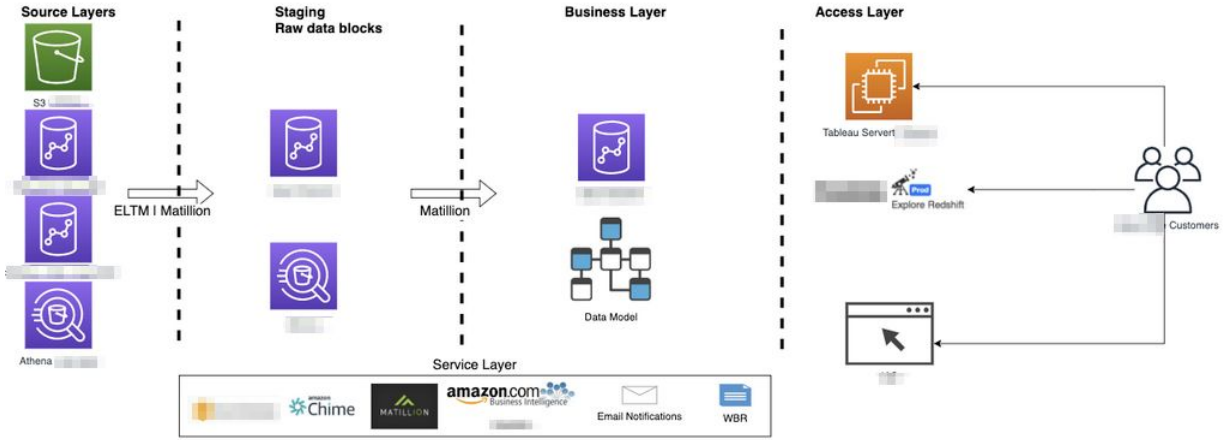
Starting state:

- Custom SQL extracts (ETL) into CSV files
- Manual combine CSV into Tableau Workbooks
- No automation and monitoring

Use cases:

- Alexa Feature Launch monitoring
- Alexa new languages launch monitoring
- Dive deep into bad NLU KPIs
- Churn Model for Alexa Customers

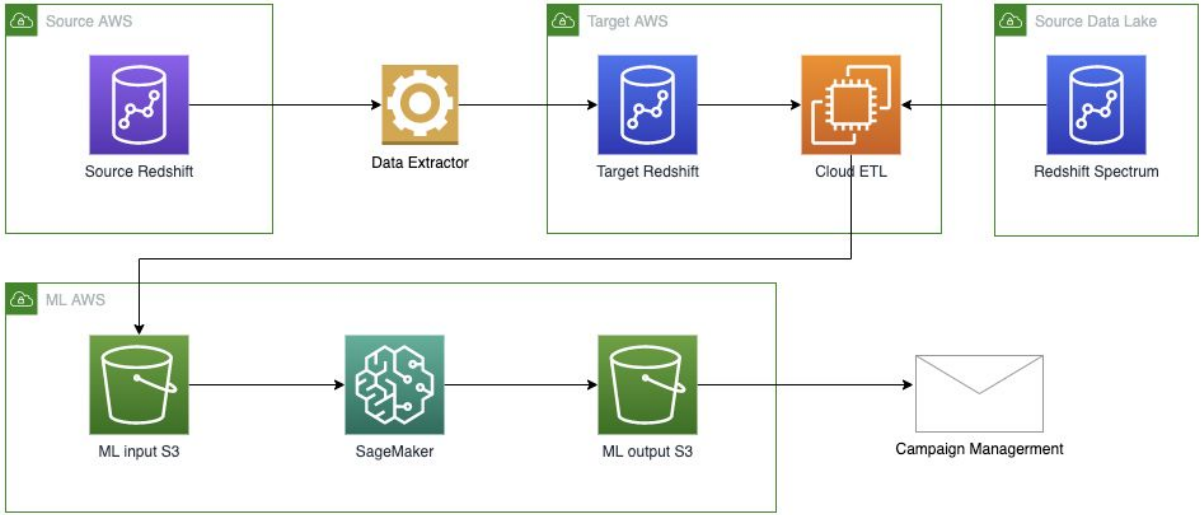
Amazon Redshift, Matillion ETL, Tableau



Solution:

- Amazon Redshift “Sputnik” (6 months before COVID)
- Tableau
- Custom SDE data product for drill down from Tableau Dashboard

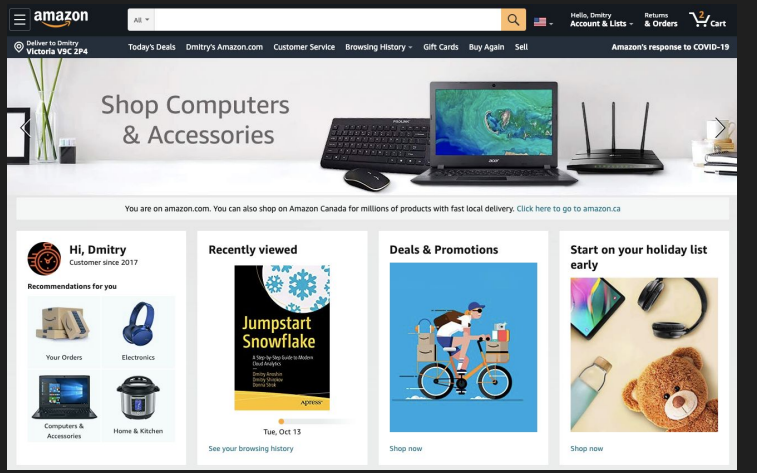
Churn model with SageMaker



Solution:

- Consolidate data from Redshift Source and S3 Data Lake (Spectrum)
- Unload data into S3 and feed to SageMaker

Case 3: Feature Engineering for ML



Company:

- Amazon Retail

Team:

- 1 Data Engineer
- 1 BI Engineer
- 3 ML Engineers
- 1 SDE
- 1 Technical Product Managers
- 1 Sr. Product Manager

Starting state:

- Only product idea and requirements

Use cases:

- Onsite Attribution
- Customer Perception based on Survey data

Machine Learning: The High-Interest Credit Card of Technical Debt

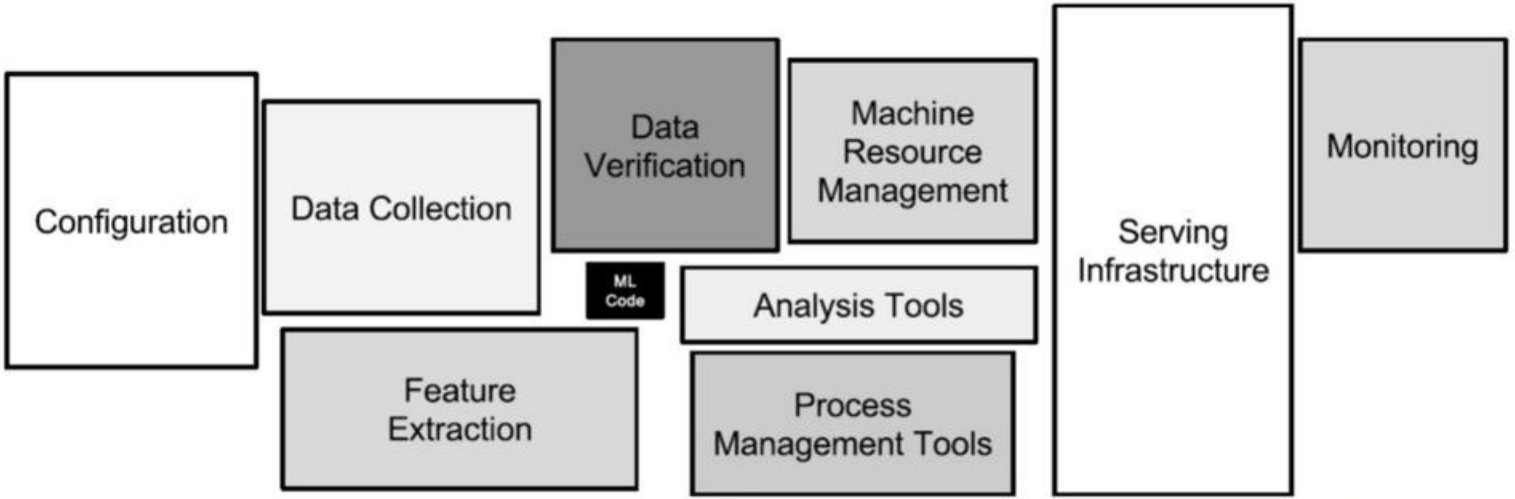
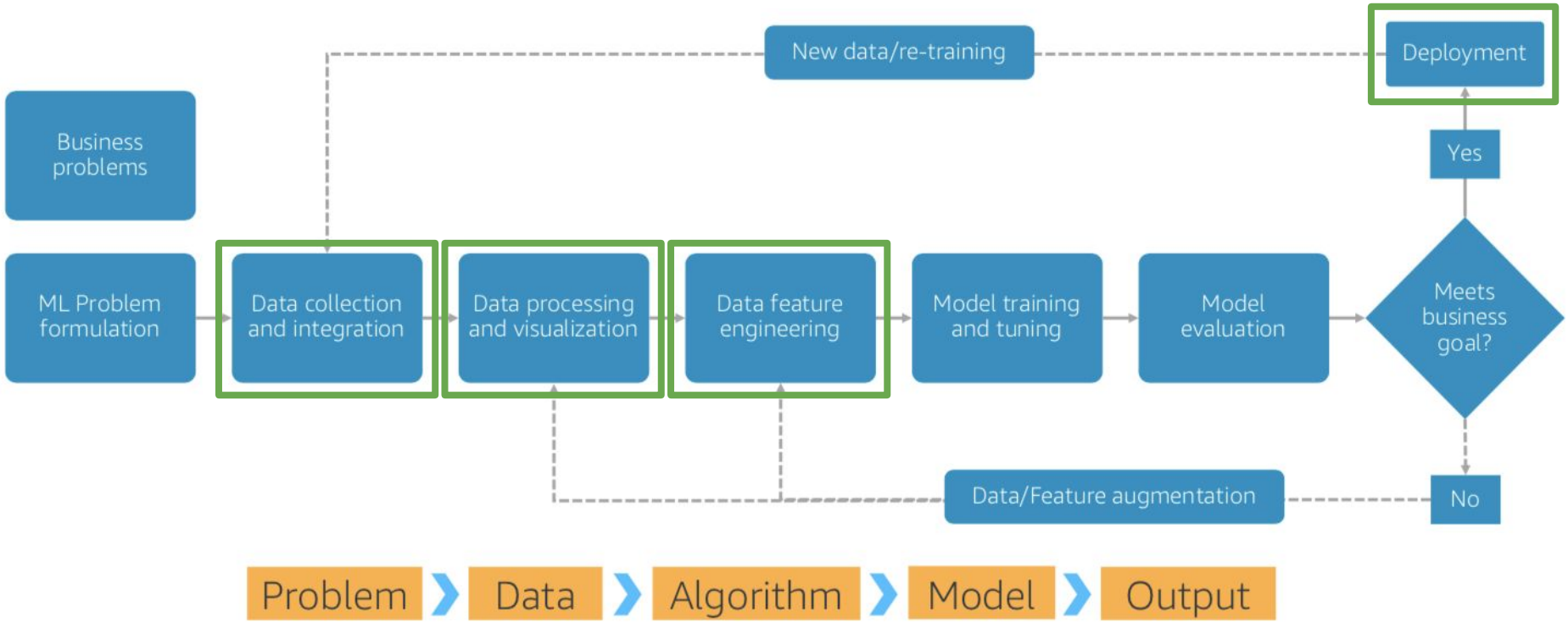


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43146.pdf

Role of Data Engineer on ML project



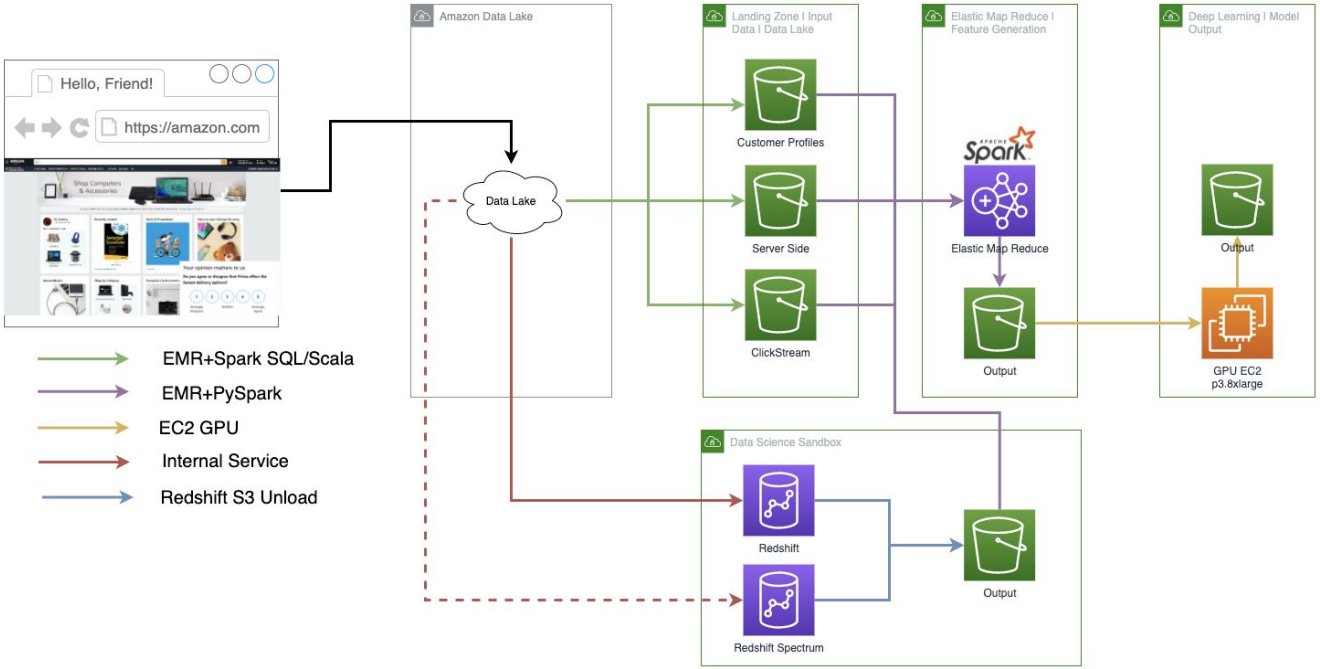
Role of Data Engineer in ML Projects

- Prepare data for model
- Data Quality
- Automation data pipelines
- Automation ML build (MLOps) -> Model freshness metric.

Don’t have:

- Typical Data Warehouse
- Datamodel (dimensional modeling and etc)
- Business Intelligence (Dashboards)

AWS Feature Store for ML



Solution:

- Extract data from Core Data Lake using in-house Spark and Scala into own my AWS account S3
- Using EMR+Spark combine the data

Case 4: Delta Lake for Gaming Analytics

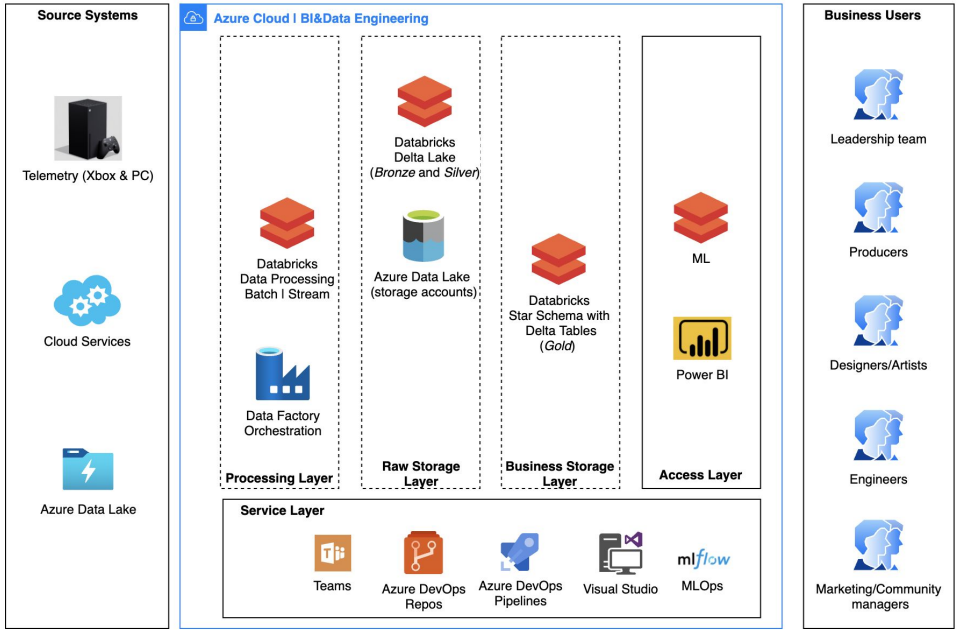
Company:
- Microsoft Xbox, The Coalition

Data Team:
- 2 Data Engineers
- 1 BI Engineer
- 2 Data Scientists
- 1 Sr. Manager Analytics

Starting state:
- HDInsights + Hive (staging)
- SQL Server + SSIS on-premise for Fact/Dim tables

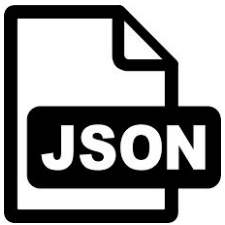
Use cases:
- Player behaviour and engagement
- Monetisation
- Game balance
- Skin/Weapons/Features use
- Multiplayer balance

Azure Databricks Delta Lake for Gaming



Solution:

- Hourly batch from Azure Storage into Delta Lake
- RAW and STAGE easily APPEND mode (could be Structured Streaming), FACT is MERGE/INSERT+DELETE
- Power BI for BI
- Azure Bicep
- Azure DevOps



Raw Tables (Bronze)
Method: Append
Trans: Minimum

Staging Tables (Silver)
Method: Append
Trans: JSON Schema

Fact Tables (Gold)
Method: Merge
Trans: Heavy

Case 5: Analytics Solution for SaaS product (pre-IPO)

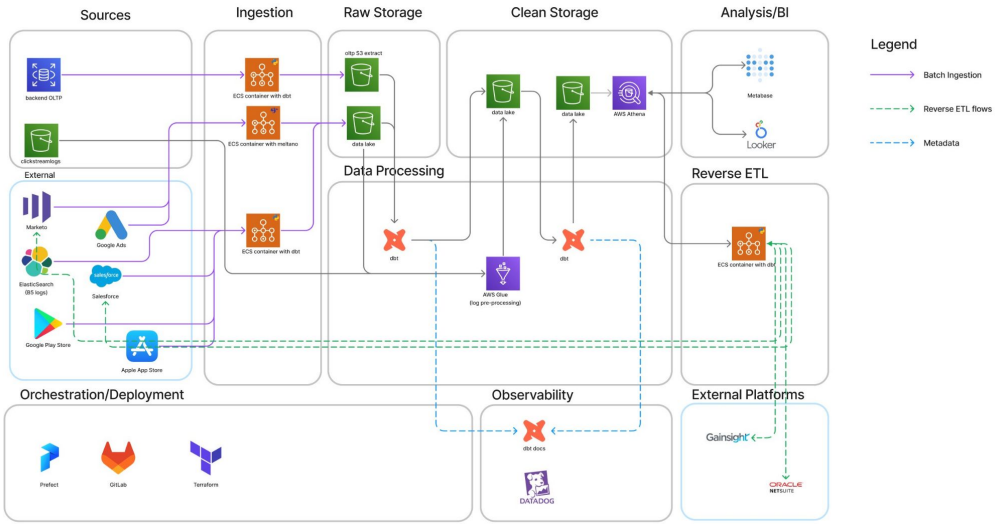
Company:
- Canadian SaaS product (pre-IPO)

Data Team:
- 6 Data Engineers
- 6 Analytics (dbt+sql+looker) Engineers
- 2 Product/2 Marketing/4 Financial
- DE/AE/Product/Marketing/Finance Managers
- 1 Sr. Manager Analytics
- 1 Director DE
- 1 VP Data

Starting point:
- Open Source aka tech debt
- dbt adapter for Athena
- Data Solution built by DevOps

Use cases:
- Financial/Marketing/Product reporting
- BI Dashboards
- Finance reporting for IPO
- Product telemetry project with Snowplow

AWS Athena Data Lake with open source ETL/ELT



Solution:

- Meltano for extractions
- dbt core for Athena and Redshift
- Athena SQL engine on top of Parquet data lake
- Looker as BI
- Metabase as Ad-Hoc SQL
- Glue Spark for Logs processing (product, clickstream)
- Terraform for DE infra
- Git, CI/CD
- Prefect for Orchestration
- ECS for running dbt, prefect agent
- AWS Batch (old scheduler)

Moving to Redshift DW to address dbt Athena constraints.

Case 6: Analytics Solution for public SaaS product company

Company:

- Gartner leader, SaaS product

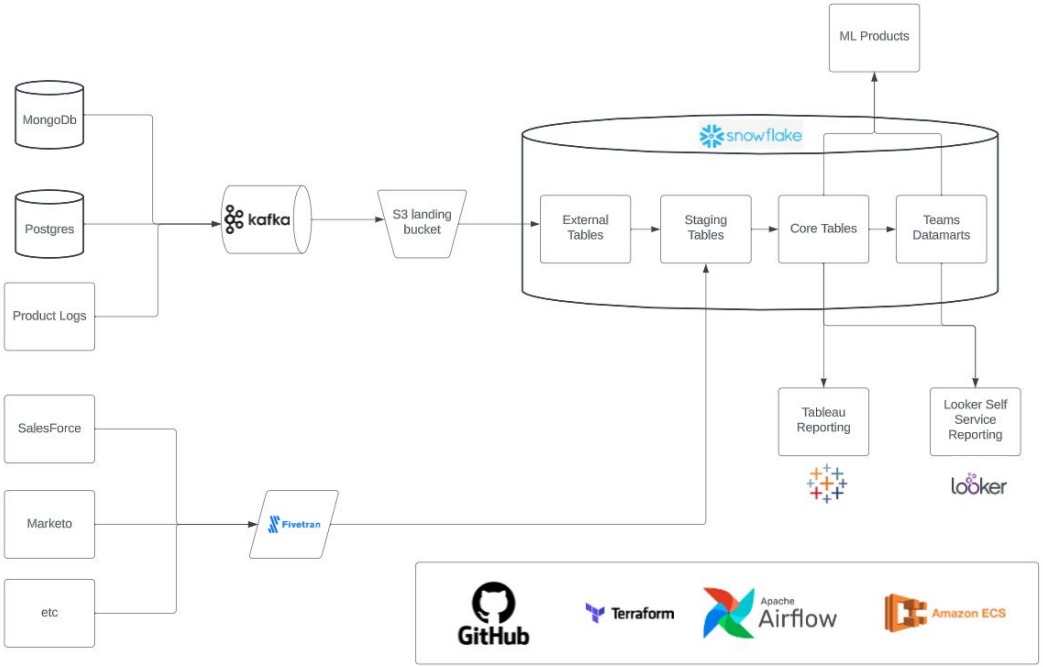
Data Team:

- Data Platform team ~8 engineers (manage Kafka stream)
- Data Pipeline team ~6 engineers (manage ingestion into Snowflake)
- Data Warehouse team ~6 engineers (manage core DW models)
- ML team
- BI team
- 1 Sr. Manager
- 1 VP of Data

Use cases:

- Data Warehouse and BI reporting in Tableau
- Self Service Reporting in Looker and Snowsight
- Customer facing data insights
- ML features for product
- Technical Sales team dashboards

Snowflake Data Warehouse



Solution:

- Snowflake is a core Data Warehouse
- dbt core for Snowflake
- Platform team run Kafka Connect for MongoDb, Application Logs and etc, write into S3.
- Snowflake External Tables for S3
- Fivetran for brining SalesForce, Marketo, Google Sheets and etc
- Looker as self-service BI
- Tableau as curated BI
- Terraform for DE infra and Snowflake
- Git, CI/CD
- Airflow for Orchestration (ideally push to Kubernetes)
- ECS for running dbt
- Cube.js for customer facing vizzes

Thinking about Snowpark, Iceberg and Dynamic Tables

Case 7: Analytics Solution for Largest Telecom

Company:

- Largest North America Telecom Company

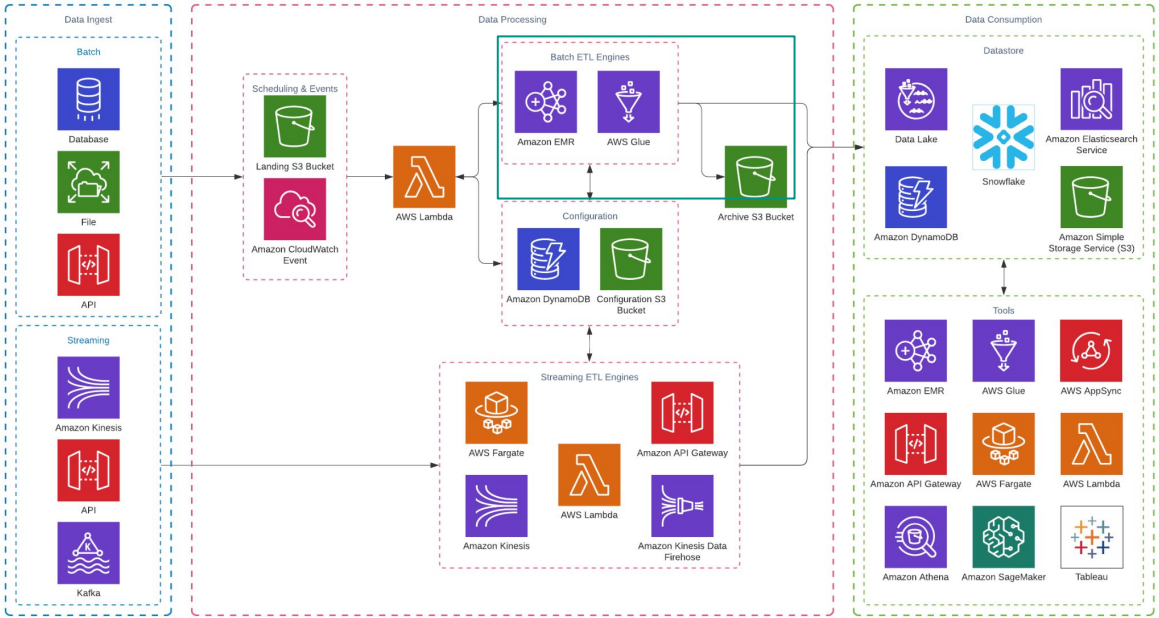
(Big) Data Team:

- 1 Data Engineer
- 2 BI Engineer
- 1 Front End Engineer
- 3 Architects
- 6 ETL (BigData) developers
- 1 BigData Manager

Use cases:

- Ingestion data into Snowflake stage
- Transform raw data in S3 data lake
- ML pipelines for data quality

AWS Serverless Data Lake on S3



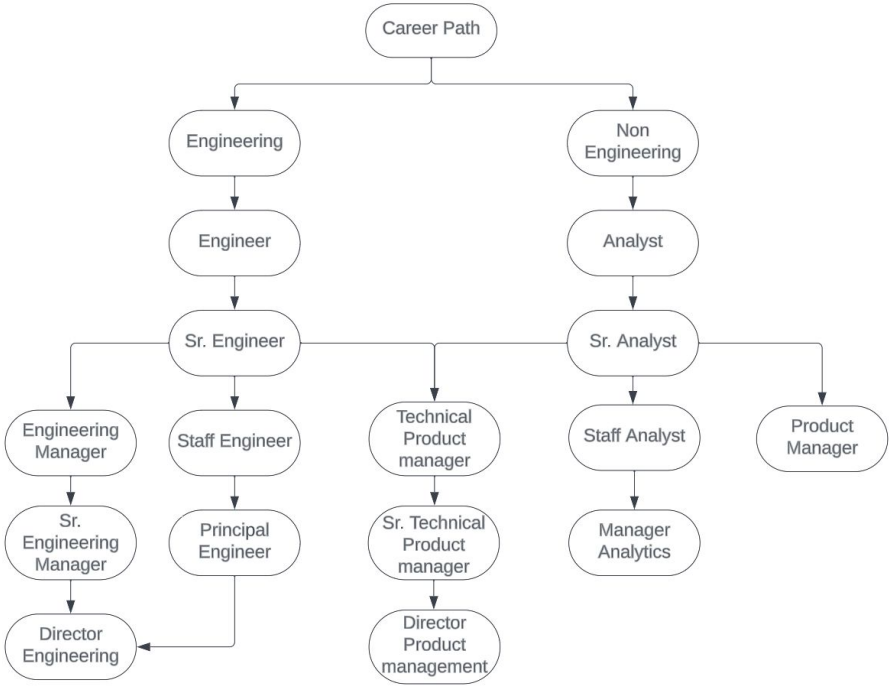
Solution:

- Git Repo
- EMR+Spark jobs
- Tigger based on file volume, process data and save to S3
- Athena for SQL
- Some jobs push data to Snowflake
- Using Tableau, Alteryx
- Each job a set of YAML configuration files
- Job create EMR cluster and Terminate when succeed.
- Sphinx for documentation and data quality portal
- Also has many more data tools

Thinking about moving all to Snowflake and consolidated multiple instances in one region. Also, short term to use EMR Serverless.

Career thoughts

Engineering or Management path for Data Analytics roles (non ML)?



Challenges:

- Compensation
- Career grow
- Layoffs waves (business is business…)
- Spending more than earn, easy access to credit money (cars, phones, houses, boats with weekly payments)
- New challenge - Mortgage rate

Solution:

- Change company for rise
- Have another source of income (contracts, freelance, youtube channel etc)
- Always monitor market and keep CV updated
- Always learn new skills

Canada and US salaries

DE in Canada:
- DE 120k - 145k CAD + bonus/stocks
- Sr. DE 150k - 180k CAD + bonus/stocks
- Staff DE 180k - 250k CAD + bonus/stocks

F(M)AANG - pay higher, but usually 1.5-2x less in US $ in compare with US.

Summary:
- Cost of living very high
- Tax is very high
- Real estate is crazy high
- Salary is low

DE in US:
- DE 120k - 150k US + bonus/stocks
- Sr. DE 140k - 200k US + bonus/stocks
- Staff DE 220 - 250k US + bonus/stocks

In US bonuses/stocks is big part of total comp.

F(M)AANG - pay higher and allow high bonuses

Summary:
- Cost of living is good (non SF, NY)
- Tax is ok (non SF, NY)
- Real estate is ok (non SF, NY)
- Salary is high

Optionally there is an hourly rate 65-100 CAD/$ per hour in case of Incorporated (contract/freelance)

Summary

- DE position is on demand
- AE position is on demand only due dbt but it is not real position
- Snowflake and Databrics are popular
- Snowflake is know as very expensive
- dbt almost every project with cloud DW
- Azure and GCP less popular
- Big companies split teams with purpose as a result lack of big picture (end to end)
- Open Source for Data Solutions by DE/AE tend to slow down the project and rise the technical debt.
- Open Source has issues and require to look into source code, modify or fix bugs, it is not about “fast time to market”
- Amazon 2 Pizza Team + Product Manager is the best way
- Data Quality still, Data Catalog, Data Governance still the issue like 20 years ago
- Lots and lots trainings and products, easy to lose focus