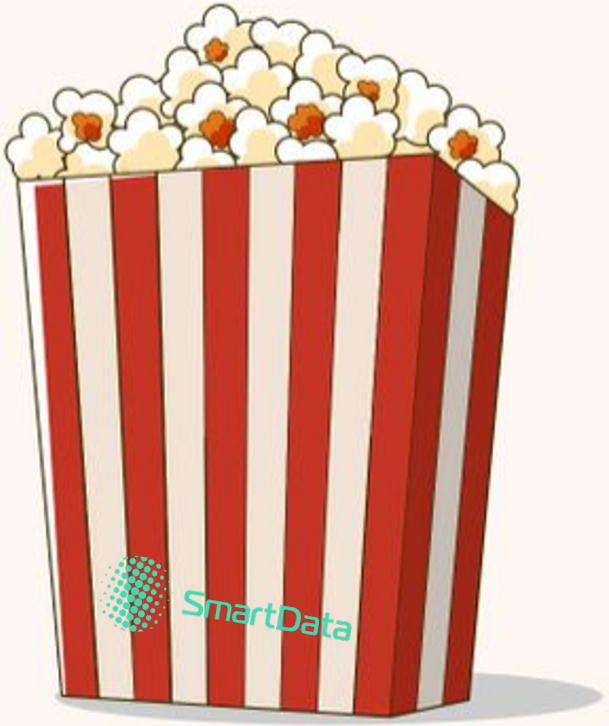
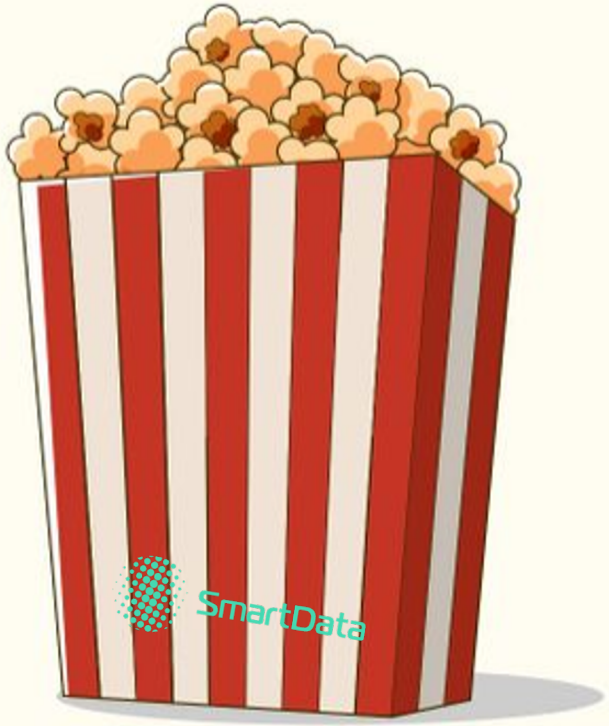
2 Types of Data Engineers

~~Sweet~~ **Gentle**



~~Salty~~ **Hardcore**



Disclaimer

All thoughts are mine and don’t represent Microsoft or any other company. Based on my experience and environment I worked over decade.

Outline

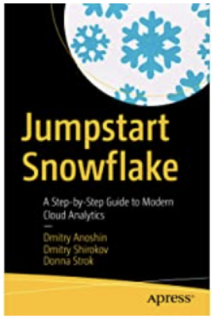
- About myself
- 2 Use Cases
- 2 DE for our use cases
- Some Architectures reviews

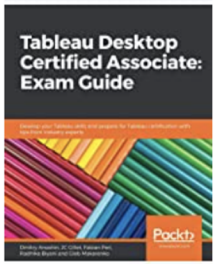
- 11+ years in Analytics
- Moscow, Montenegro, Winnipeg, Vancouver, Victoria, Seattle, Boston
- 5 years @Amazon, now @Microsoft Gaming
- Tableau, Snowflake, Microsoft, AWS user groups and meetups

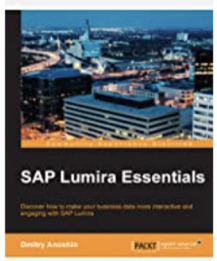
DataLearn.ru
4000 Students
- DE 101
- DS&ML 101
- SQL 101

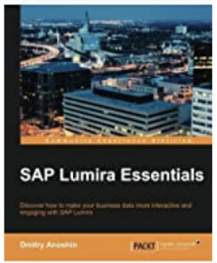
University of Victoria

Инжиниринг Данных
@rockyourdata
Technologies









Use Cases

Use Case 1 - Online Store 

Marusya is running online store of hunting rifles.

Running on premise

Use Case 2 - Mobile Games 

Innokenty is creating mobile games and publishing them to Google and Apple stores.

Running on public cloud

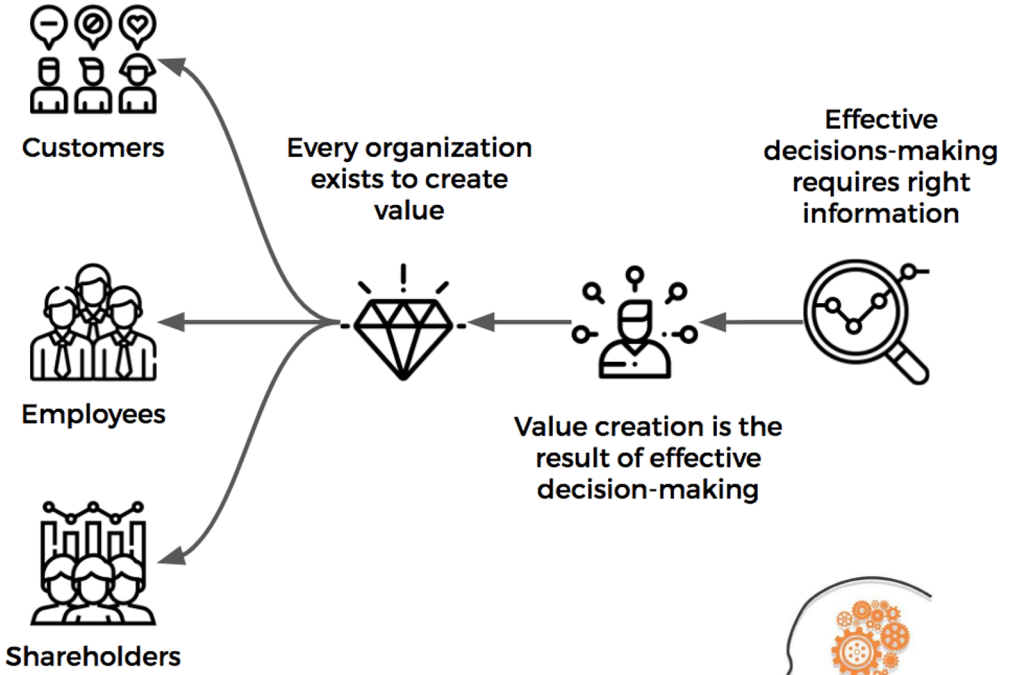
<u>Their Goals:</u>

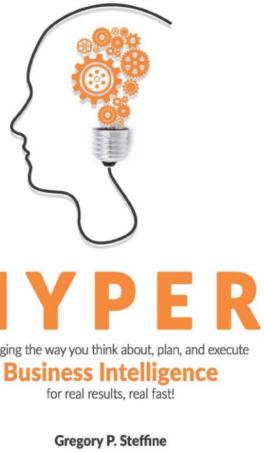
- Measure business performance and KPIs
- Handle scale of successful business and survive during New Years holidays peaks
- Identify areas of improvement and business optimization
- Increase customer experience and decrease costs of running business

What do they need?

What is Analytics?

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





Use Cases

Use Case 1 - Online Store 

Marusya is running online store of hunting rifles.

Running on premise

Use Case 2 - Mobile Games 

Innokenty is creating mobile games and publishing them to Google and Apple stores.

Running on public cloud

Their Goals:

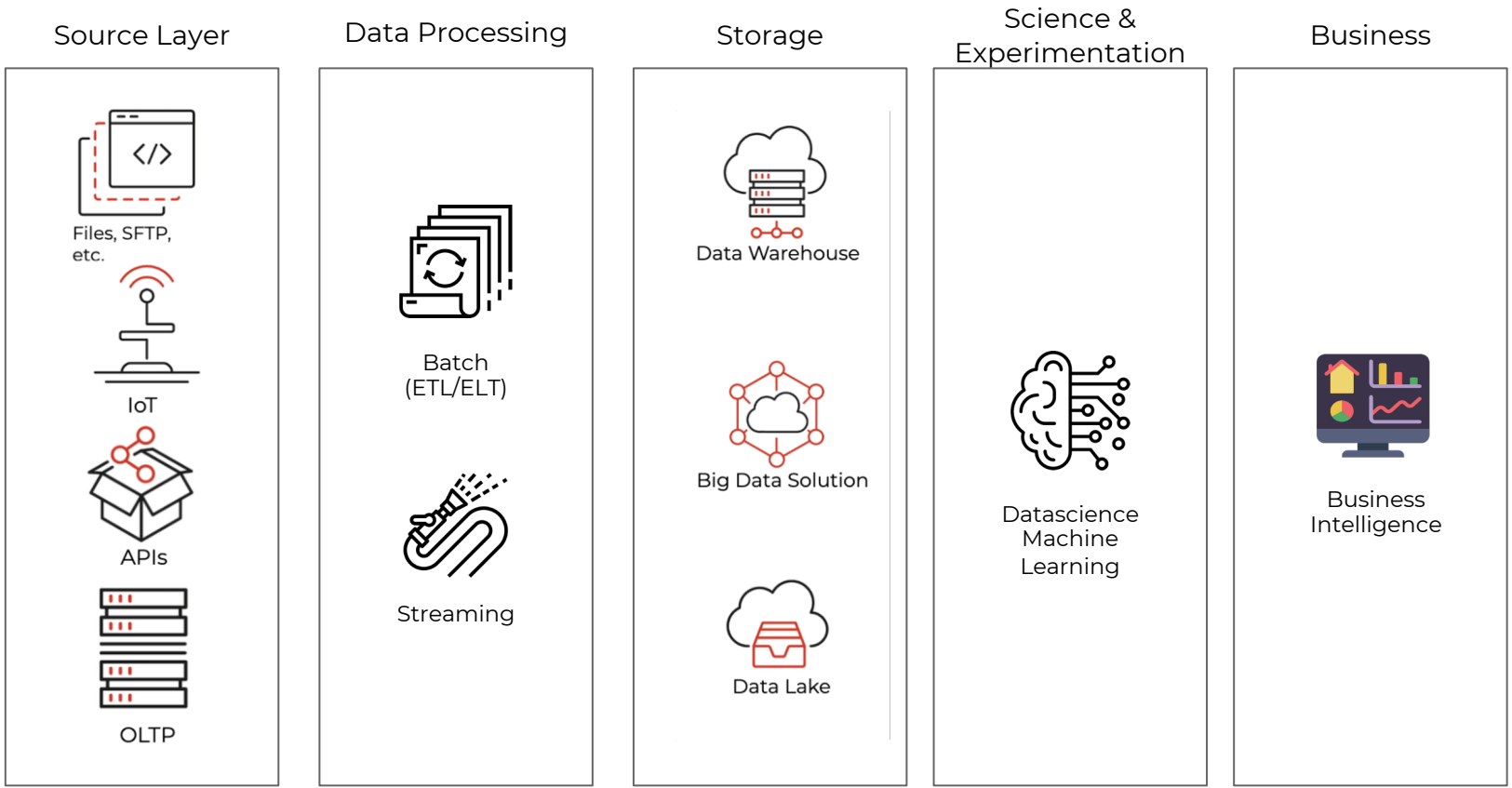
- Measure business performance and KPIs
- Handle scale of successful business and survive during New Years holidays peaks
- Identify areas of improvement and business optimization
- Increase customer experience and decrease costs of running business

What do they need?

How to reach the goal (from analytics perspective)

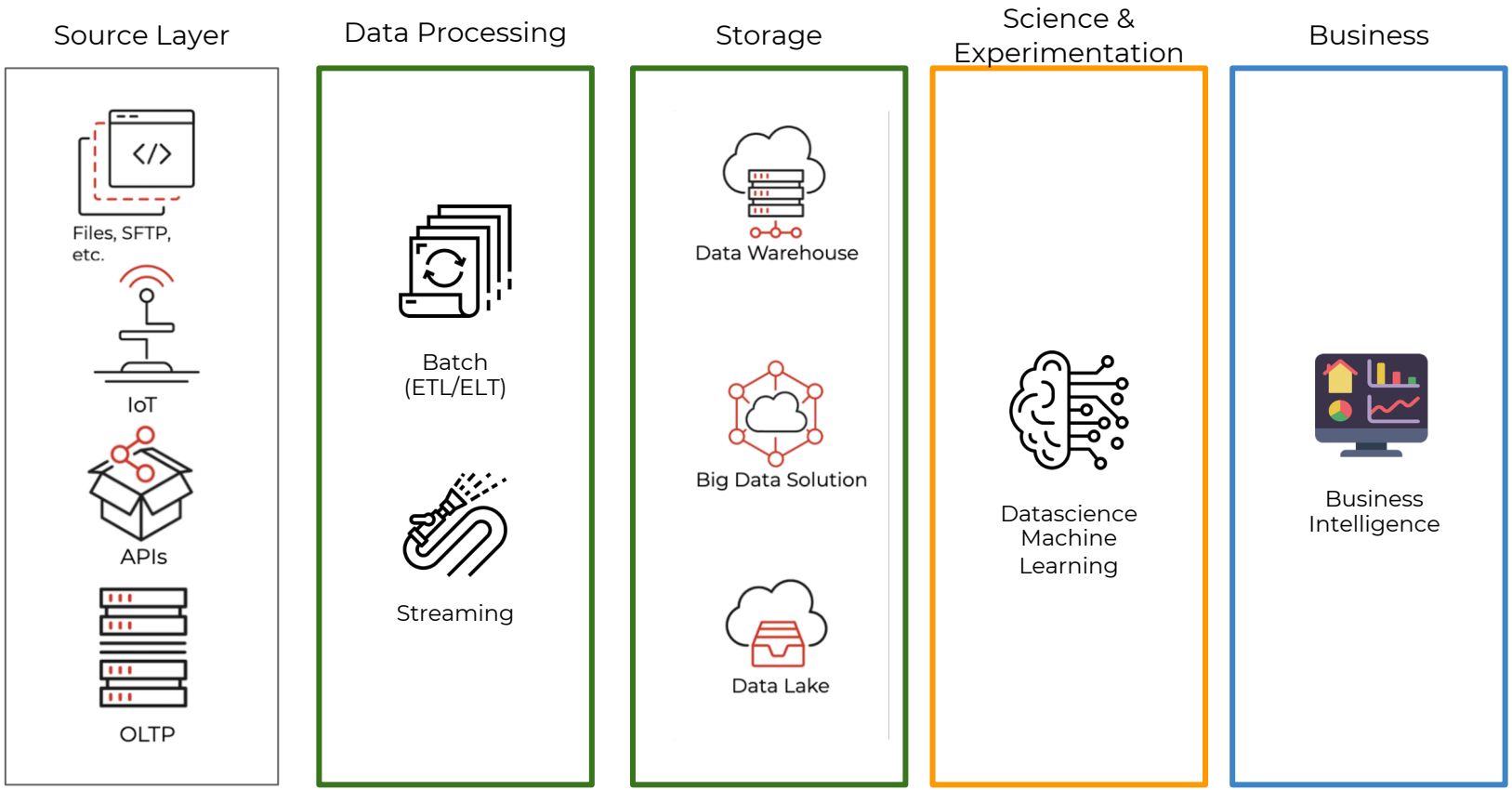
1. (Modern) Data Stack
2. Data Team
3. Budget
4. Data and Hypothesis

Modern Data Stack



Modern Key Layers and roles

(1)Product Manager - manage data product.



(2)Data Engineer

(3) ML Engineer
Data Scientist

(2)BI Engineer

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.

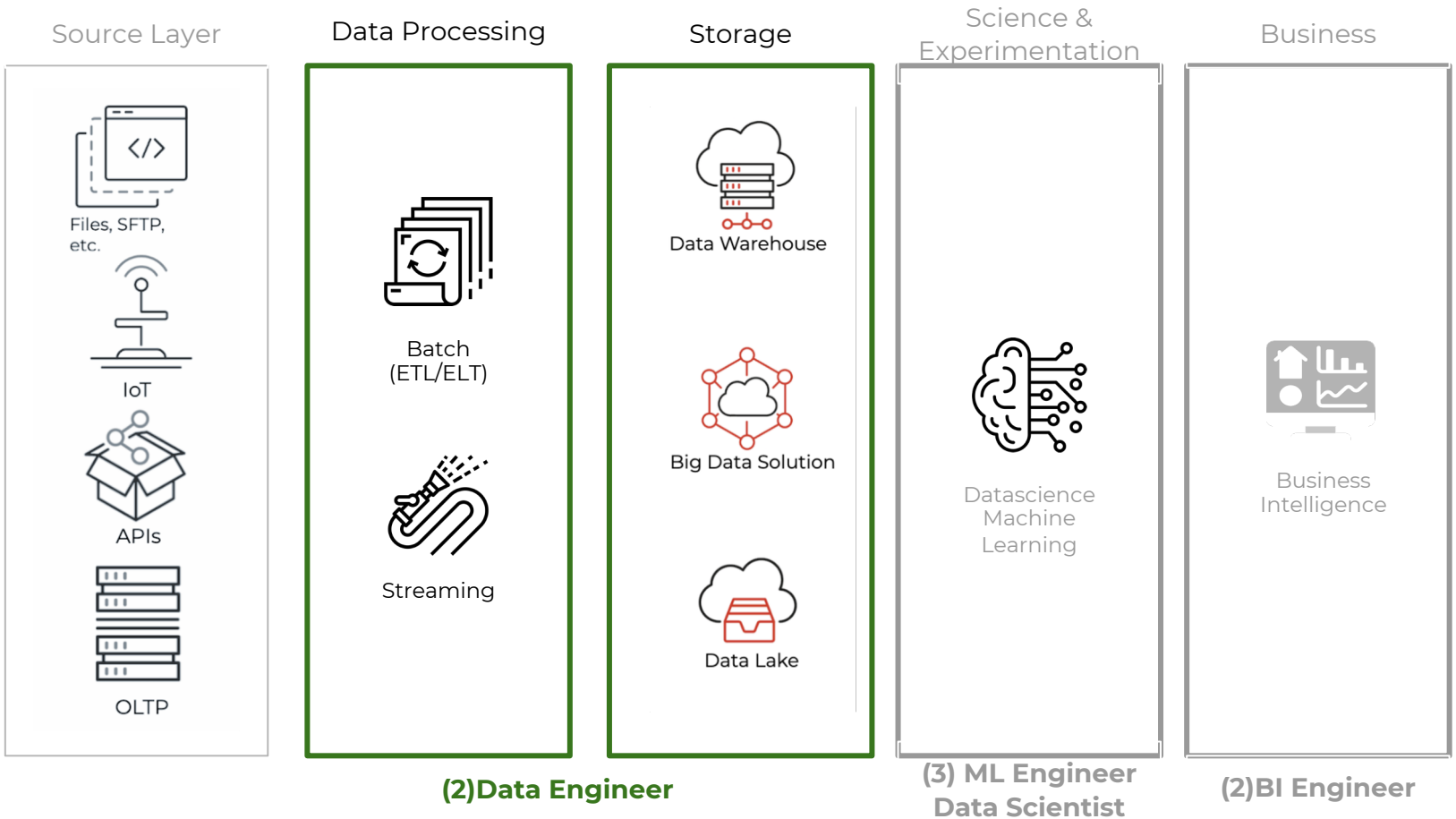
Amazon: Data Engineers tackle some of the most complex challenges in large-scale computing. Most of the work they do involves storing and providing access to data in efficient ways.

TL;DR:

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

DE Key Layers

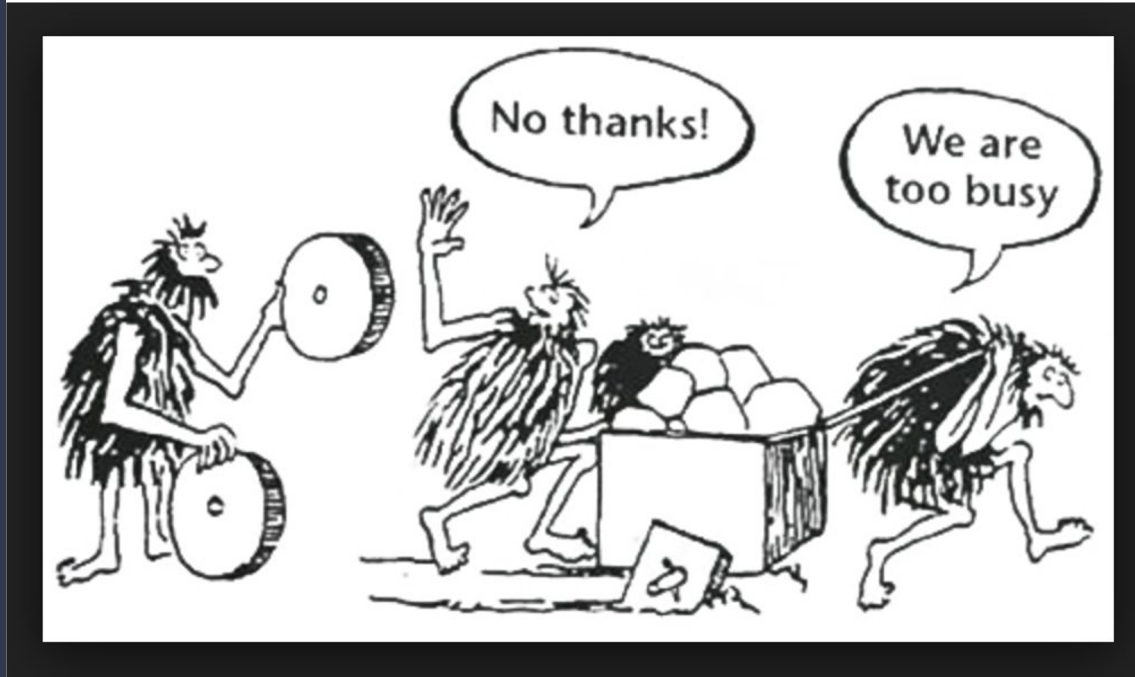
(1)Product Manager - manage data product.



How would you do this for our friends?

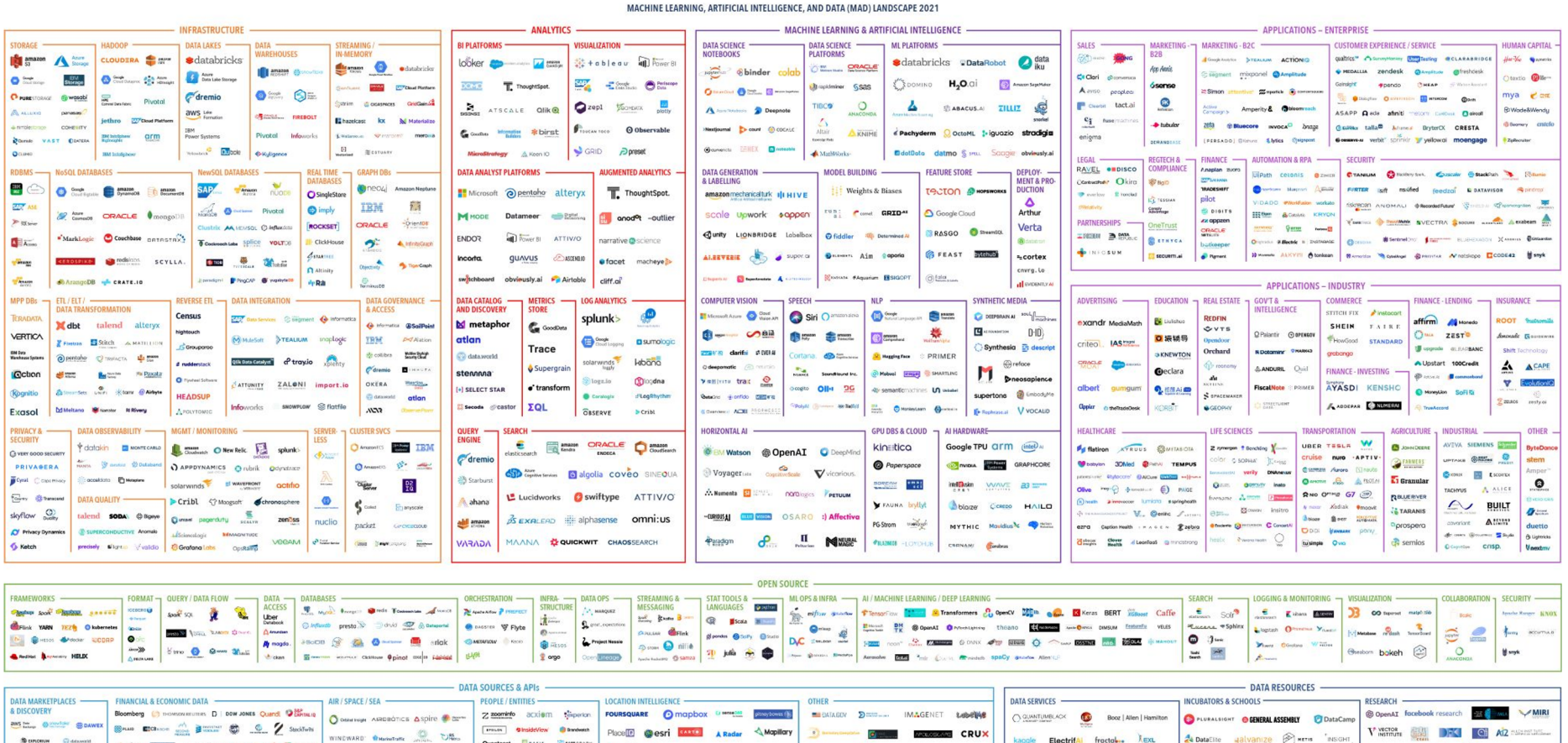
This is the question!

Data Engineer should know the answer!



With DE you can - “move fast, break things”(c)...

Data and AI Landscape 2021



FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Till this moment we talked only about abstract DE....

Use Cases

Use Case 1 - Online Store 

Marusya has Computer Science degree and short on budget.

Code and Open source

Use Case 2 - Mobile Games 

Innokenty has Skolkovo MBA and has funds from father of his wife.

Applications with UI and Support

<u>Their Goals for Data Engineering:</u>

- Hire Data Engineer
- Consolidate data into single storage
- Make data accessible by BI and DS
- (Hopefully) Security and Compliance with Privacy
- (Ideally) Documentations, DevOps

New DE, welcome to the team!

With help of Data Engineering telegram channel @rockyourdata ;) they hired DE.

Marusya found Anna.

Ana was a software engineer and worked with Java and Go. She did coding and built backend services.

Innokenty found Valera.

Valera was a data analyst and supported business with data insights by building data solutions from scratch.

docker

ClickHouse

plotly | Dash

aws

looker

snowflake

DataRobot

Fivetran

dbt_

Anna's Architecture

Source Layer	Data Processing	Storage	Science & Experimentation	Business
Files, SFTP, etc.	Batch	ClickHouse	Datascience Machine Learning	Apache Superset
IoT	Streaming	Hadoop		plotly | Dash
APIs				Business Intelligence
OLTP				

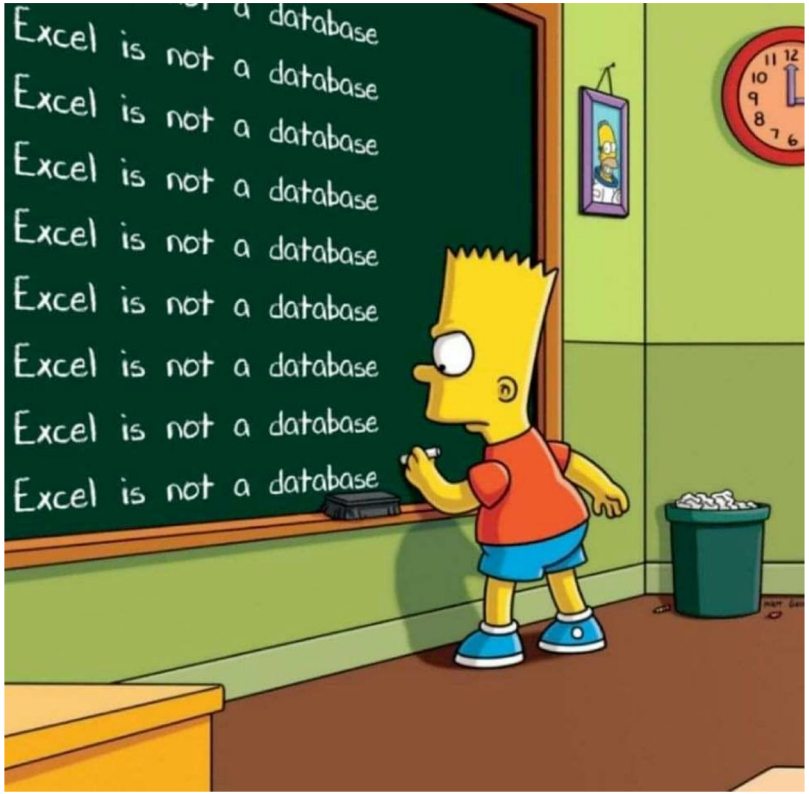
Valera's Architecture

Source Layer	Data Processing	Storage	Science & Experimentation	Business
Files, SFTP, etc. IoT APIs OLTP	dbt_ Fivetran Batch (ETL/ELT) Streaming with Snowpipe	snowflake	alteryx DataRobot Datascience Machine Learning	looker Business Intelligence

Just in case: Valera's on-premise Architecture

Source Layer	Data Processing	Storage	Science & Experimentation	Business
Files, SFTP, etc. Microsoft Dynamics OLTP	Microsoft SQL Server SSIS	Microsoft SQL Server		Microsoft SQL Server Reporting Services Microsoft SQL Server Analysis Services

Just a reminder for Valera;)

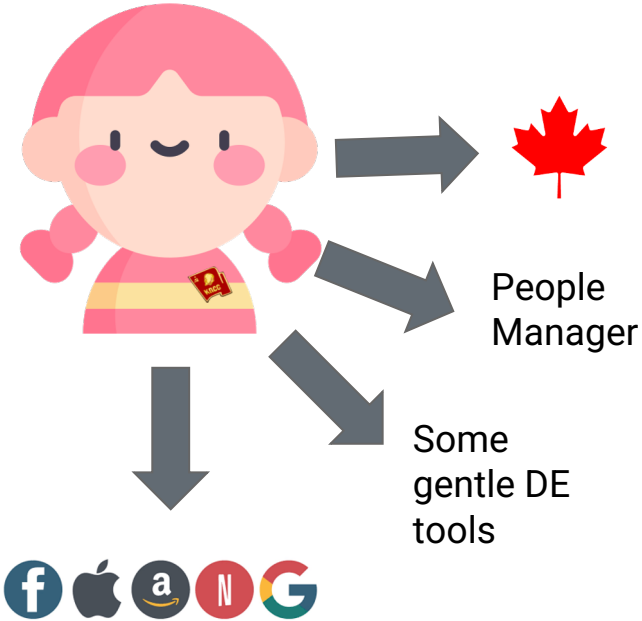


Outcome from 1 to 5 (5 is the best)*

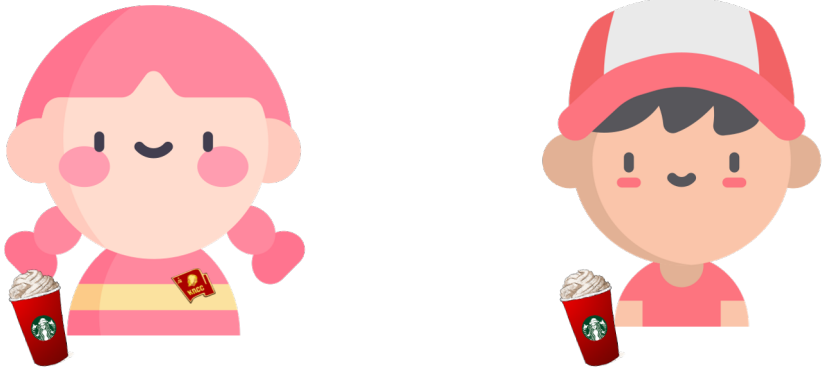
	 Avg. 3.25	 Avg. 3.25
time to market*	2	5
maintenance	3	4
engineering excellence	5	3
CI/CD	5	0
Dev/Prod	5	2
time to onboard new empl	2 (longer)	4 (faster)
easy to replace?	1 (could be hard)	4 (relatively easy)
easy of scale	3	4
is “CEO” happy?	Absolutely!	

* with time to provision hardware

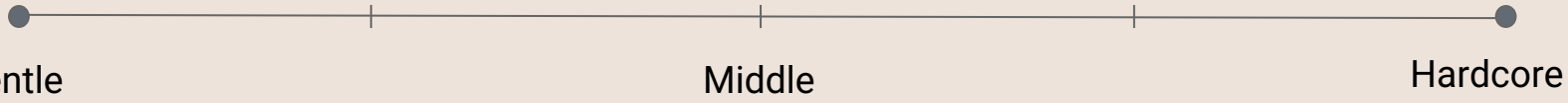
Where to Move?



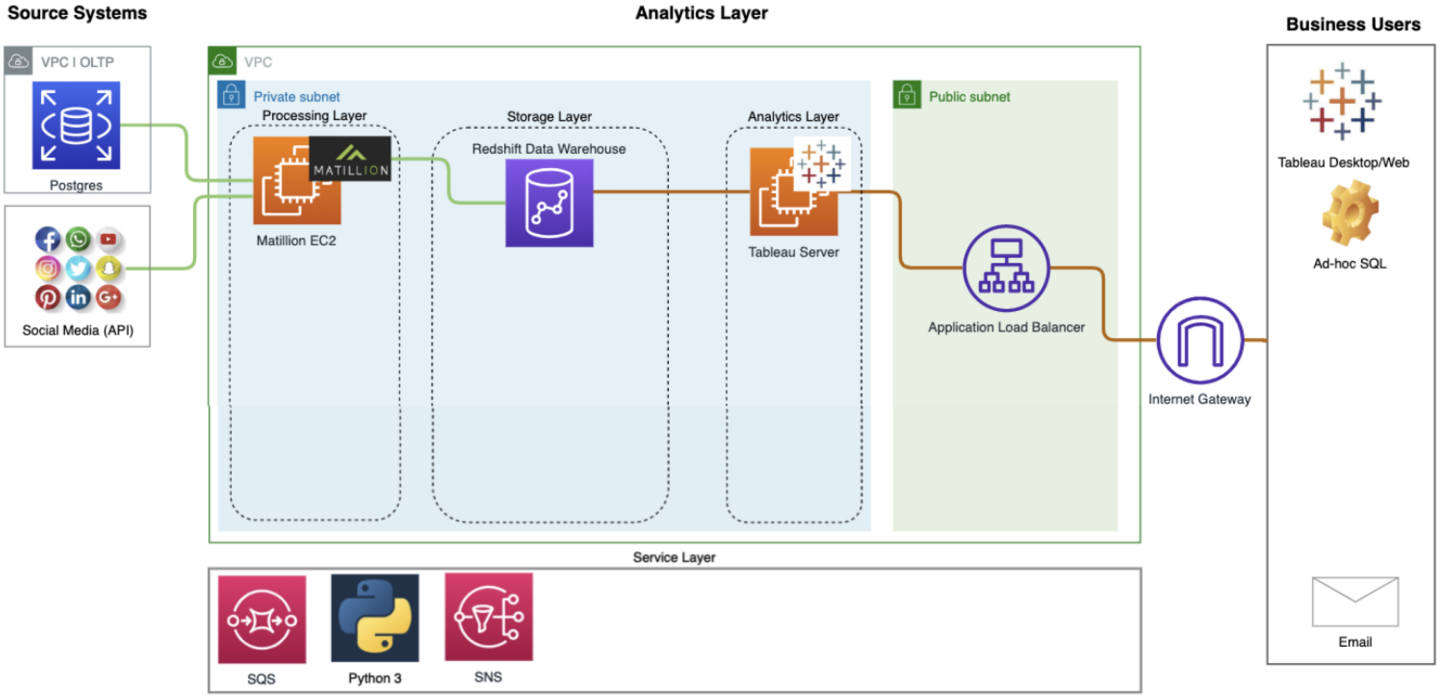
Better -> Work in a team!



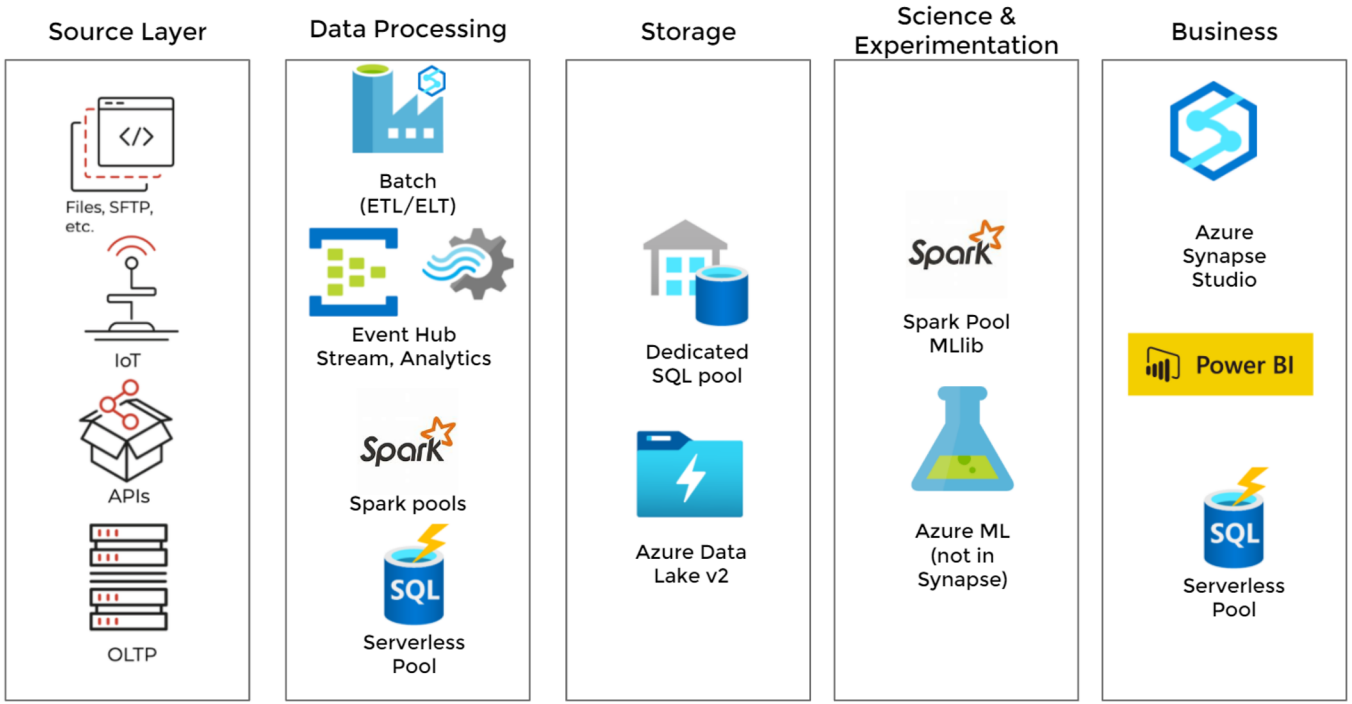
Let’s check come solutions on *Gentle to Harcore scale*



AWS Data Warehouse

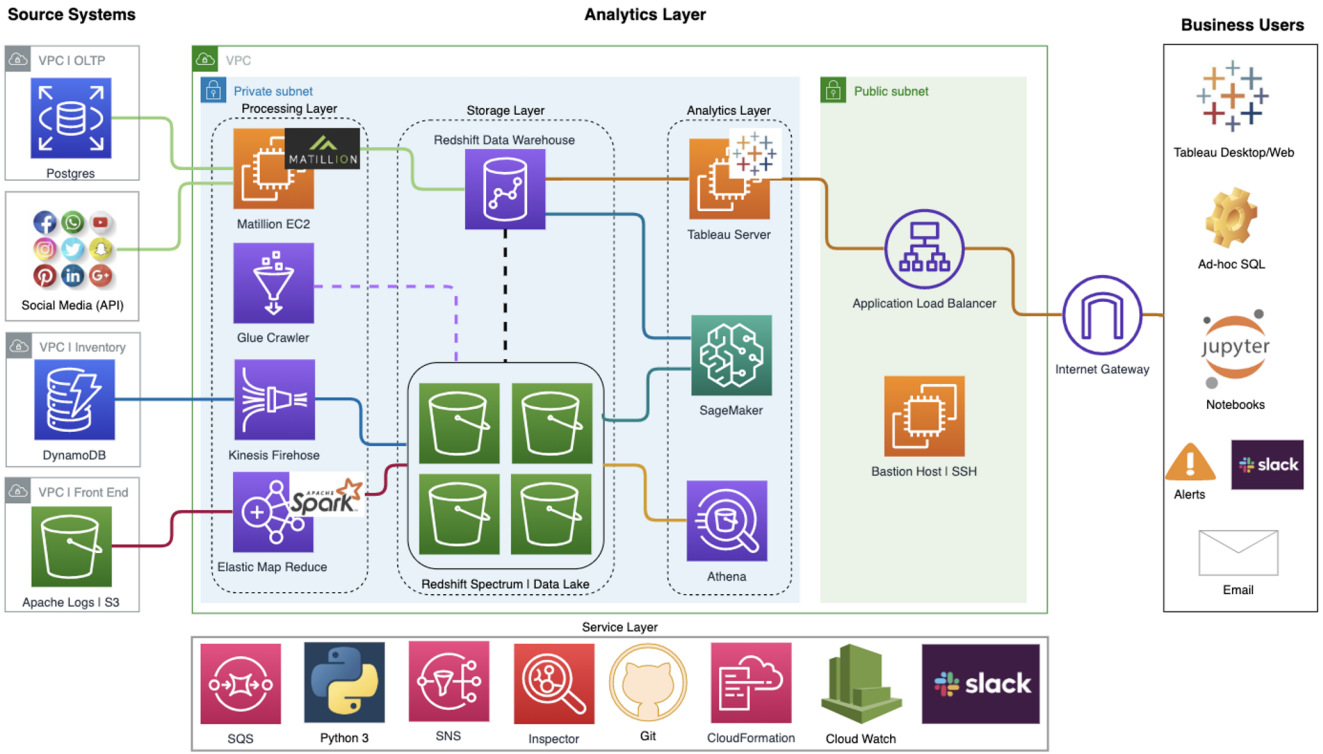


Modern Solution with Synapse

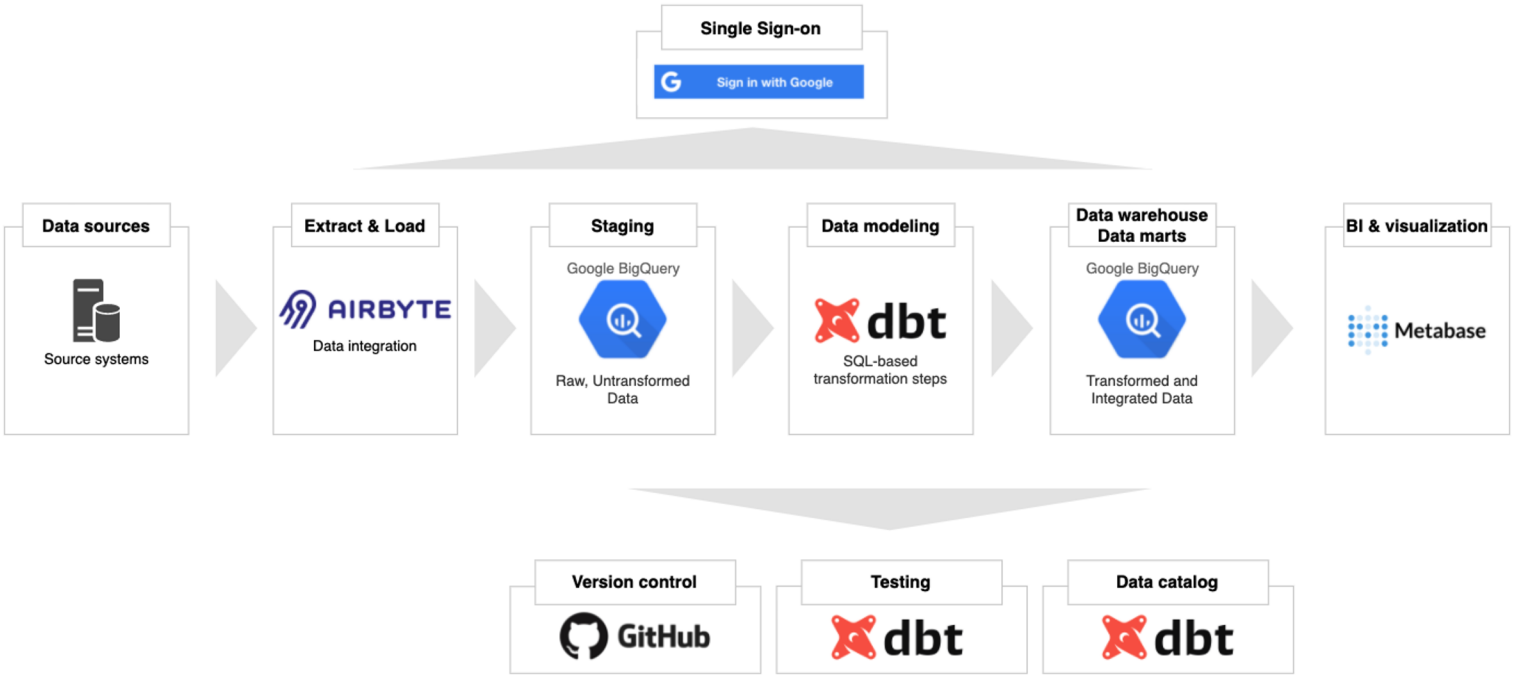


AWS Modern Data Platform

Gentle — Middle — Hardcore

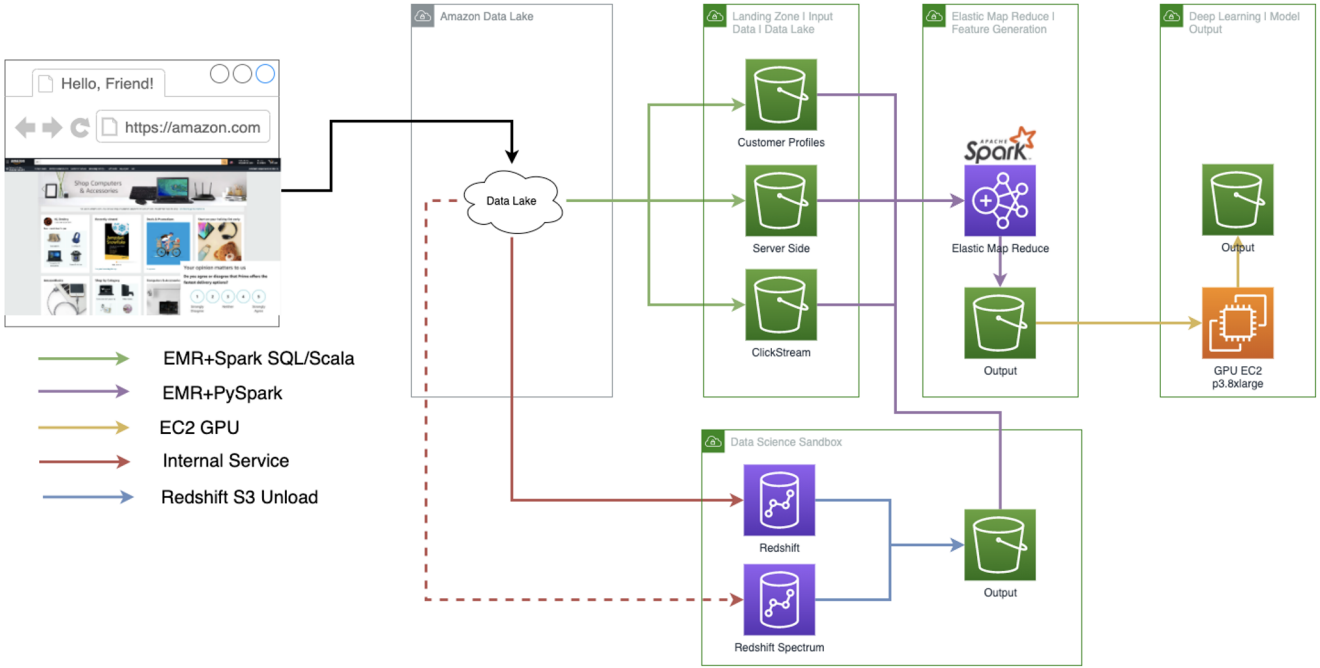


Example of Modern Data Stack

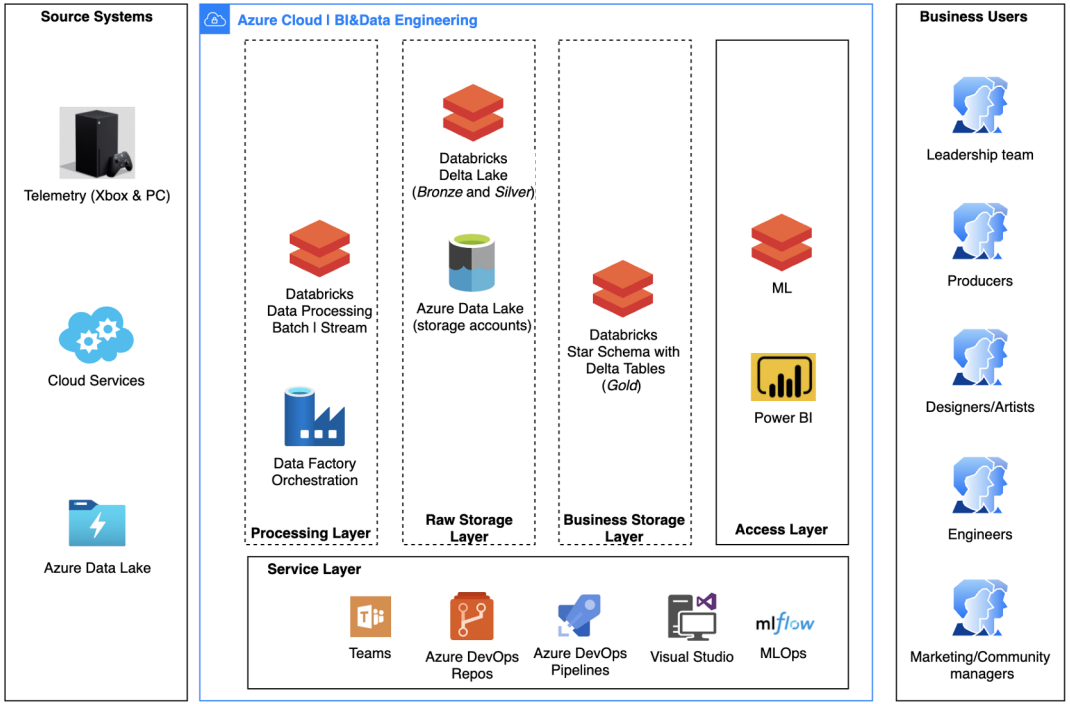


https://towardsdatascience.com/bootstrap-a-modern-data-stack-in-5-minutes-with-terraform-32342ee10e79

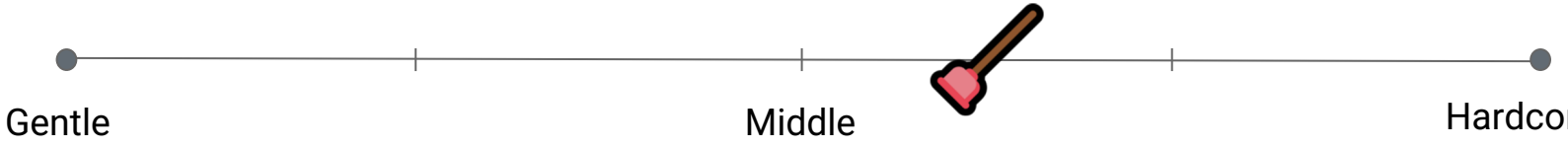
AWS Feature Store for ML



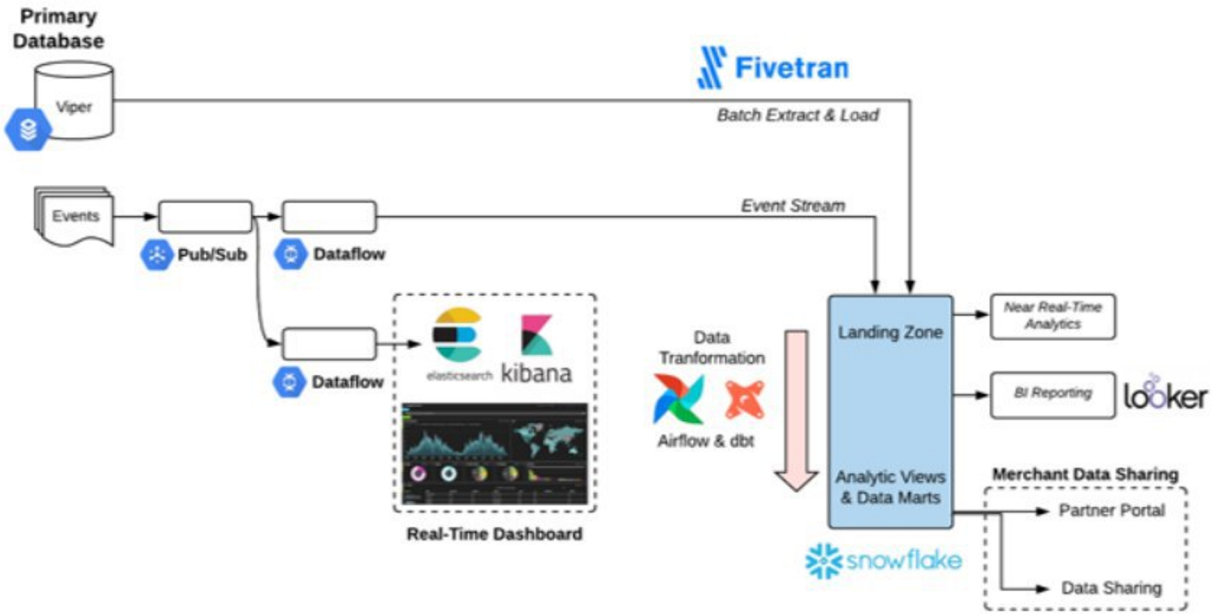
Azure Delta Lake for Gaming



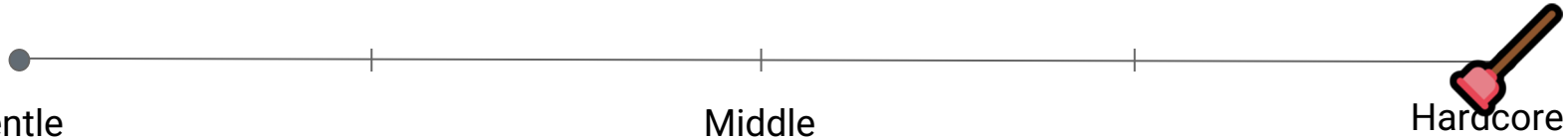
Solution with Fivetran, Airflow and DBT

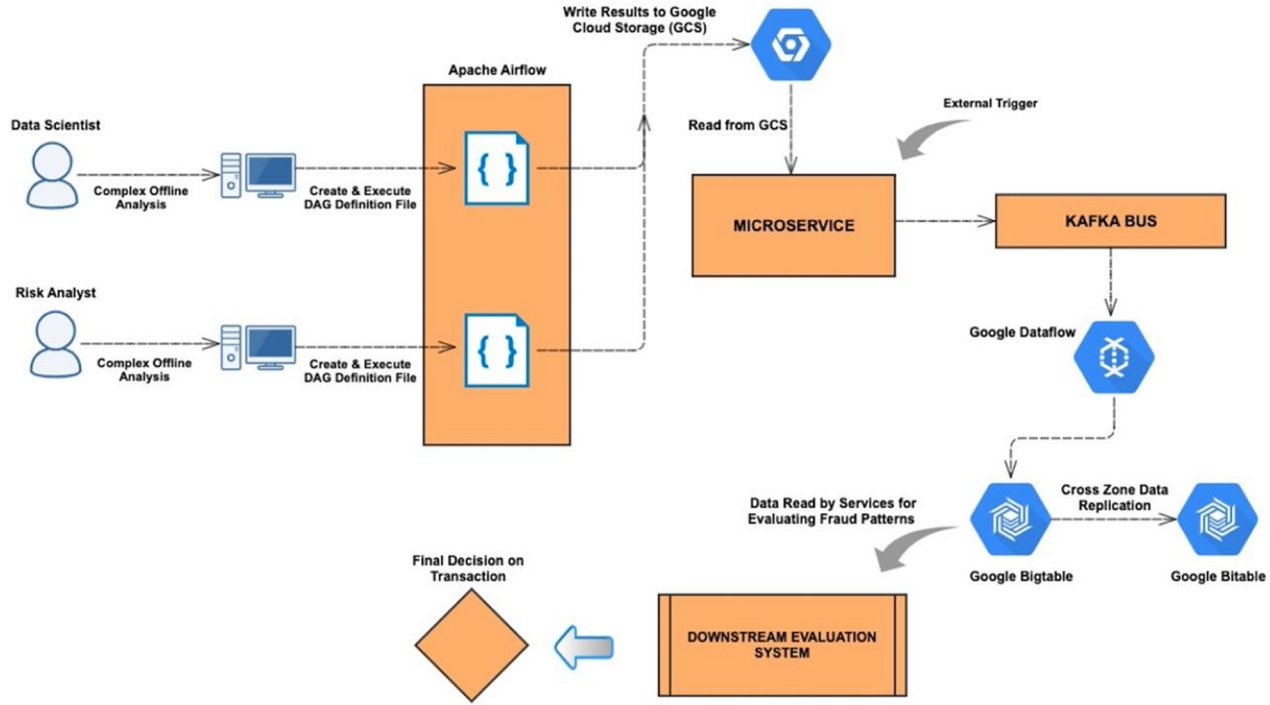


Target State



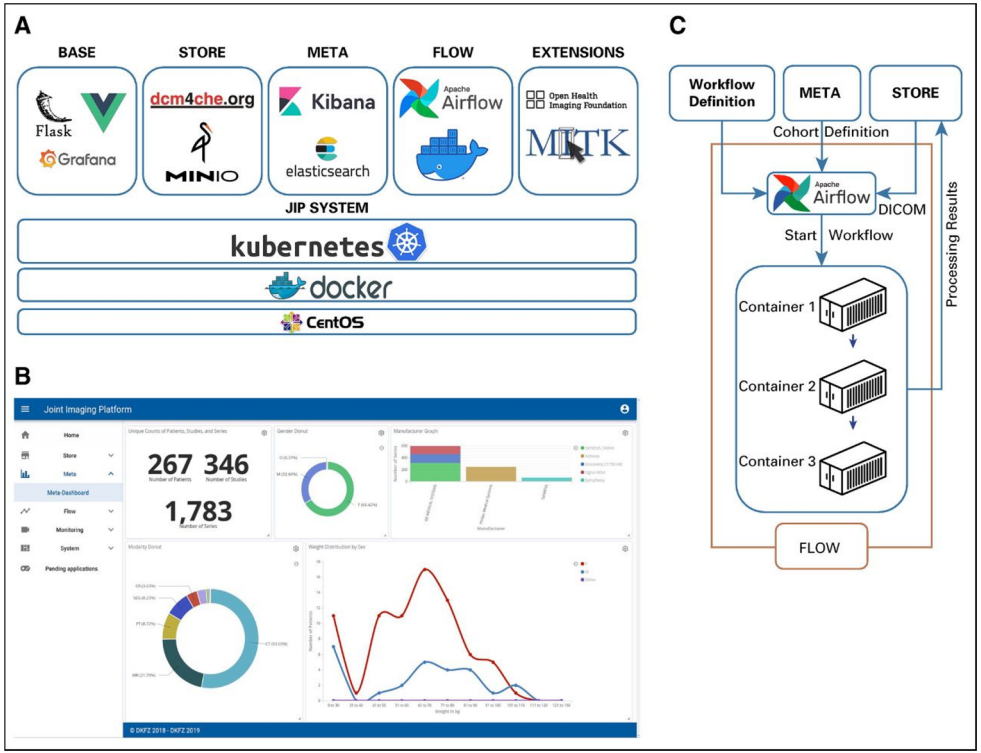
Data Pipeline at WePay





Clinical Data Analytics (Open Source)

Gentle — Middle — Hardcore



https://ascopubs.org/doi/10.1200/CCI.20.00045

Summary

- Focus on business value and desired outcome
- Help gentle people to pick up gardcore skills, everyone is learning
- Help hardcore people to get business knowledge and work with diverse team
- There is no bad solutions, there are bad implementations
- and have a common sense

For everything else - @rockyourdata