Modern Data Stack

A primer

Agenda

- Disclaimers
- History
- A tale of two data architectures
- Birth of modern data analytics
- A silent revolution
- Modern data stack and beyond

Disclaimers

- The term “Modern Data stack” is very subjective. Just like “Big data” or “Agile” or “Renaissance”
- I have my own biases.
- I have over-simplified many concepts in this presentation.
- We won’t cover everything
  - Real time Analytics
  - ML and Data Science
  - Containerization and Kubernetes

1990s

TDK
IEC I/TYPE I NORMAL POSITION
D90
B
TDK
D
LUCKY
TYPE I (NORMAL) POSITION/NORMAL
BIAS 120μs EQ
Super EF

OLTP vs OLAP

- Online Transaction Processing (OLTP)
  - Relational databases like MySQL or Oracle
- Online Analytical Processing (OLAP)
  - Big vertical Analytics Databases often provided by IBM or Oracle
  - Similar to OLTP databases mostly in features and limitations.
- SQL (and Excel) galore

2005+

11:45 AM
MENU

Web Scale

- Google and Yahoo want to index the entire internet
- Traditional databases can not manage this scale
- Hadoop is born in 2005 at Yahoo (Google’s MapReduce algorithm)
  - Lot of small, cheap machines working together to run computations.
  - Distributed processing of big-data

2010+

iPhone
8GB
Designed by Apple in California Assembled in China Model No. A1203
iPhone
8GB

Obsession with Scale

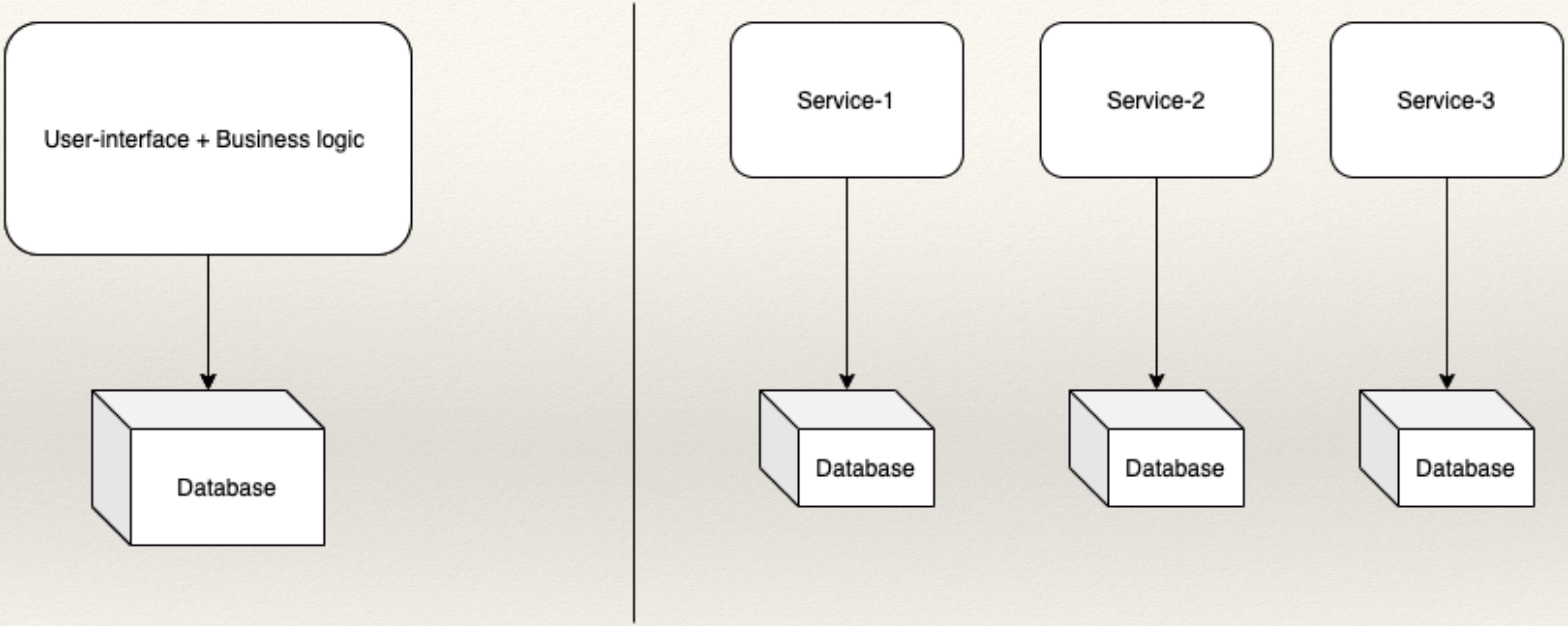
- More and more data being collected
- Everyone trying to harness “scale”
- Hadoop is the de-facto big-data processing engine
- Data-lake emerges as a viable architecture

Trending…

54

Rise of the cloud

Microservices and NoSQL



User-interface + Business logic

Database

Service-1

Service-2

Service-3

Database

Database

Database

Summary so far

- Before 2000 —— OLAP databases
- 2000+ —— Web2.0, Hadoop
- 2010+ —— Hadoop ecosystem matures, Datalake
- 2010+ —— Cloud becomes de-facto, Microservices

DataLake

Data Lake

- Dump all data into a central lake
- Typically on Hadoop or Amazon S3
- Cheap
- File formats like Parquet (columnar storage + compressed)
- Processed using Hadoop or Apache Spark (NoSQL)
- Often Schema-less
- Sloooooooooooooooooow

2012–2020

XI
I
X
THU 3
II
8
4
7
5

2012-2016

- Cloud adoption shoots through the roof (50-60% of all enterprise workload)
- Managed services and Serverless Compute revolutionise how we write software.
- Redshift is launched in 2012. First enterprise grade cloud data warehouse
- SQL renaissance
- ETL —> ELT
- Birth of modern analytics

2016-2020 - A silent revolution

- Organizations fail to get a good ROI on their data investments
- Data Quality, Management and Governance become centerstage
- Productionizing and Operationalizing data become the new challenge
- Call for Data democracy (DataMesh)
- DataOps and Software engineering practices adopted by data community
- Emergence of new roles - **Analytics Engineer, Data Product Owner, Data Governance Manager**
- Birth of the Modern Data Stack

Cloud
DataWarehouse

Data Warehouse

- Alternative to the Data Lake architecture
- Fully managed by a cloud provider
- Extremely fast at petabyte scale
- Fully serverless (scaling becomes a cost issue)
- SQL everywhere

Summary so far

- Before 2000 —— OLAP databases
- 2000+ —— Web2.0, Hadoop
- 2010+ —— Hadoop ecosystem matures, Datalake
- 2010+ —— Cloud becomes de-facto, Micro-services
- SQL renaissance
- Serverless Cloud Data warehouses emerge as an alternative to DataLake
- Data Management and Ops are becoming difficult to manage
- Modern Data Stack

Modern Data Stack

…highly specialized tools **come together** to form the modern data stack, a scalable, low barrier to entry group of technologies that startups and enterprises alike can adopt to **drive immense value** from their data…

https://snowplowanalytics.com/blog/2021/05/12/modern-data-stack/

…businesses encounter many challenges and complexity in leveraging and operationalising their data assets. Driven by the scalability and cost-effectiveness of cloud data warehouses/lakes, the modern data stack is a suite of tools and patterns that have emerged to address these challenges and lower the barrier for data integration…

https://www.thoughtworks.com/en-de/insights/blog/data-engineering/meet-the-modern-data-stack

Features

- Cloud based
- Low barrier to entry
- Agile, Democratic
- Pay-as-you-go
- Integrate with well-known dwh/lake solutions
- SaaS model (Commercial Open source)
- Highly integrated with other tools in the tool chain
- Serverless and low maintenance (Scaling is a billing problem)
- SQL oriented

Categories of tools

- Data ingestion
  - Behavioural data ingestion (**Snowplow, Mixpanel, Google Analytics**)
  - Transactional data ingestion (**Fivetran, Airbytes, Meltano, Singer**)
- Storage
  - Cloud data warehouse or data lake or… Lakehouse 😠!!! (**Bigquery, Snowflake**…)
- Workflow Orchestration (**Airflow, Argo, Prefect, Dagster**…)
- Data Processing and Transformation (dbt, Spark, Flink, Presto)

Categories of tools…

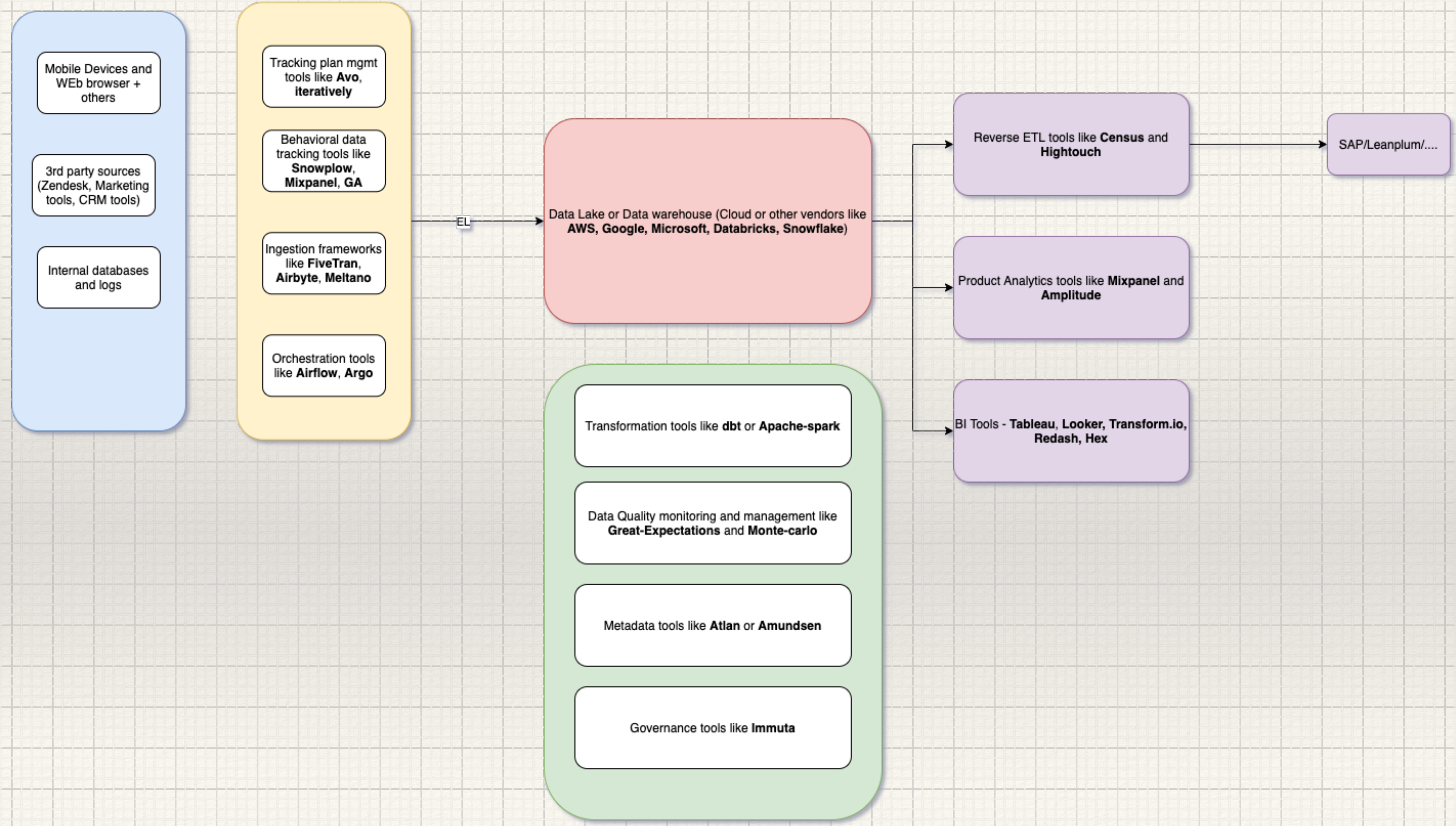
- **Data Management**
  - Data Quality (**Great-expectations, Monte-carlo, Soda.io**…)
  - Data cataloging and metadata (**Atlan, Amundsen** and many more…)
  - Data lineage (**Amundsen, Marquez, Atlan**…
  - Governance, security, access control (**Immuta, Ranger**…)

Categories of tools…

- Data Analysis
  - Product Intelligence (**Mixpanel, Amplitude**)
  - Business Intelligence (**Looker, Tableau, PowerBI, Redash** and many more)
  - Notebook style tools (**Hex** and many more)
  - Headless BI (**Transform.io, Metriql, Minerva**)

Categories of tools…

- Data Operations
  - Reverse ETL (**Hightouch, Census**)
  - CDP systems (**Segment, mParticle**)
- AL and ML tooling



Mobile Devices and WEb browser + others

3rd party sources (Zendesk, Marketing tools, CRM tools)

Internal databases and logs

Tracking plan mgmt tools like **Avo**, **iteratively**

Behavioral data tracking tools like **Snowplow**, **Mixpanel**, **GA**

Ingestion frameworks like **FiveTran**, **Airbyte**, **Meltano**

Orchestration tools like **Airflow**, **Argo**

EL

Data Lake or Data warehouse (Cloud or other vendors like **AWS, Google, Microsoft, Databricks, Snowflake**)

Transformation tools like **dbt** or **Apache-spark**

Data Quality monitoring and management like **Great-Expectations** and **Monte-carlo**

Metadata tools like **Atlan** or **Amundsen**

Governance tools like **Immuta**

Reverse ETL tools like **Census** and **Hightouch**

SAP/Leanplum/....

Product Analytics tools like **Mixpanel** and **Amplitude**

BI Tools - **Tableau**, **Looker**, **Transform.io**, **Redash**, **Hex**

Key Takeaways

- SQL is here to stay
- Data lake and warehouses are slowly merging
- Cloud services and serverless data warehouses enable better ROI
- Data Operations and Management is the new frontier
- Analytics Engineering will become a key role in all data teams
- Sprawl of so many tools is a concern - http://46eybw2v1nh52oe80d3bi91u-wpengine.netdna-ssl.com/wp-content/uploads/2021/10/2021-ML-AI-Data-Landscape-V2.pdf
- Headless BI and Reverse ETL will probably see more innovation

end

Attributions

- Photo of Cassette tape by Fernando Lavin
- Photo of iPod - Original Photograph - AquaStreakImage Cleanup - Rugby471, CC BY-SA 3.0, via Wikimedia Commons
- Photo of iPhone 3g - Dan Taylor from London, UK, CC BY 2.0, via Wikimedia Commons
- Photo of Apple Watch 4th Gen - Janothan Parker, CC BY-SA 4.0, via Wikimedia Commons
- Photo of Maligne Lake, Canada by Nathan Farrish
- Photo of 1200 Getty Center Dr, LA by Damon Lam
- Photo of chemical library of Lederle Laboratory by National Cancer Institute Archives
- Presentation created using Apple Keynote Theme