Sign up to Medium and get an extra one free month of Medium for new first time Members. Become a member

Open Source Data Engineering Landscape 2024

Alexey Ananyev · Following
Front-eager · Back-it

Follow 1K+ 1 Listen Share More

The original post was published on Practical Data Engineering Substack

Introduction

While the widespread hype surrounding Generative AI and ChatGPT took the tech world by storm, 2023 witnessed yet another exciting and vibrant year in the data engineering landscape, steadily grown more diverse and sophisticated, with continuous innovation and evolution across all tiers of the analytical hierarchy.

With the continued proliferation of open source tools, frameworks, and solutions, the options available to data engineers have multiplied to such rapidly changing landscapes, the importance of staying abreast of the latest technologies and trends cannot be overstated. The ability to choose the right tool for the right job is a crucial skill, ensuring efficiency and relevance in the face of evolving data engineering challenges.

Having closely followed data engineering trends in my role as a senior data engineer and consultant, I'll like to present the open source data engineering landscape for the beginning of 2024. This includes technologies in active projects and prominent tools, empowering readers to make informed decisions when navigating this dynamic technological landscape.

Why Present Another Landscape?

Why make the effort to present another data landscape? There are similar periodic reports such as the Sensor Tower **MAD Landscape**, **State of Data Engineering** and **Bessemer Open Source Top 25**. However the landscape I'm presenting is focused solely on open source tools mainly applicable to data platforms and data engineering lifecycle.

The MAD Landscape provides a very comprehensive view of all tools and services for machine learning, AI and Data, including both commercial and open source, while the landscape presented here provides a more comprehensive view of active open source projects in the Data part of MAD. Other reports such as Bessemer Open Source Top 25 and Datastack focus more on the popularity of open source projects, with a focus on the open source projects themselves, rather than the final solution.

Annual reports and surveys such as **GitHub's state of open source**, **Stackoverflow annual survey** and **DB-Engines trends** are also great sources for gauging the current trends being used or trending in the community, but they only cover limited sections (such as databases and language) of the overall data landscape.

Therefore due to my interest in open source data stacks, I've compiled the open source tools and services in data engineering ecosystem.

As without further data, here is the 2024 Open Source Data Engineering Ecosystem:

2024 Data Engineering Landscape 2024

Tool Selection Criteria

The available open source projects for each category are obviously vast, making it impractical to include every tool and service in the picture. Therefore I have followed the following criteria for selecting the tools for each category:

- Any retired, archived and abandoned projects are excluded. Some notable retired projects are **Apache Sqoop**, **Serdise** and **Apache Spire** which might still be used in some production environments.
- Projects which have been completely inactive on Github over the past year are excluded. Some notable inactive projects are included. Notable examples are **Spark Flint** and **Apache Gobblin** projects.
- Projects which are still quite new and have not gained much traction in terms of Github stars, forks, as well as blog posts, talks and conference mentions in the online communities, are excluded. Moreover some promising projects such as **OneTable** which has made some initial traction and are implemented on the foundation of existing technologies are mentioned.
- Data lakes, ML and AI tools are excluded, except for ML platform and orchestration tools. The only focusing is what's relevant to data engineering discipline.
- Different types of storage systems such as relational OLTP and embedded database systems are listed. This is because data engineering discipline involves dealing with many different internal and external storage systems used in applications and operational systems (RDS), even if they are not part of the analytics stack.
- The category names are chosen as generic as possible based on where the tool fits in the data stack. For storage systems, main database model supported is specified (e.g. OLTP, Time Series, Document, Graph, Search). For streaming, the category is listed as "Streaming SQL DBMS" also referred as as HTAP or scalable SQL database in the market.
- Some tools could belong to more than one category. **SingleStore** is both an in-memory database and distributed SQL DBMS. But I have tried to place them in the category by which they are mostly recognized in the market.
- For certain database systems, there may be a blurry line regarding the category they actually belong to. For example RisingWave claims to be a data warehousing solution, but it built on top of **Clickhouse** which is considered a Real-time OLAP engine. Therefore it is still quite ambiguous if it has the ability to support sub-second operational OLAP queries or not.
- Not all the listed projects are fully **Portable** open source tools. Some of the projects are rather **Open Core** open source. In open core models, not all components of the full system, as offered by the main hosted provider, are made open source. However, although not adopting an open source model, it is important to consider how portable and truly open source the projects are.

Overview of Tool Categories

In the following sections each category is briefly discussed.

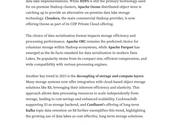
1. Storage Systems

Storage systems are the largest category in the presented landscape, primarily due to the recent surge of specialized database systems. Two lines trending categories are **vector** and **streaming** databases. **Materialize** and **RisingWave** are examples of open source streaming database systems. Streaming SQL DBMS is a relatively new category, but it is growing very fast. I have placed vector storage systems in the ML/AI platform section since they are primarily used in ML and AI stacks. Distributed file systems and object stores are also placed their own related category, that is Data Lake Platform.

As mentioned in the selection criteria section, storage systems are grouped and labeled based on the main database model and workload. At the highest level, storage systems can be classified into three main classes: **OLTP, OLAP,** and **HTAP**. They can be further categorized based on SQL vs NoSQL for OLTP engines, and Offline (non real-time) vs Real-time (sub-second) for OLAP engines, as shown in the following figure.

2. Data Lake Platform

Data Lake platform has continued to mature in the past year, and Gartner has placed Data Lake in the slope of enlightenment in its 2023 edition of Hype Cycle for Data Management.



Gartner

For storage layer, **distributed file systems** and **object stores** are still the main technologies serving as the bedrock for both on-premise and cloud based data lake implementations. While **HDFS** is still the primary technology used for on-premise Hadoop clusters, **Apache Ozone** distributed object store is catching up to provide an alternative on-premise data lake storage compatible with HDFS. Cloud providers continue to refine their own offerings for scale and speed of the Cloud Object Stores.

The choice of data serialization format impacts storage efficiency and processing performance. **Apache ORC** remains the preferred choice for optimizing storage and query performance in Hadoop ecosystems, maintaining strength as the de-facto standard for data serialization in modern Data Lakes. Its popularity stems from its compact size, efficient compression, and wide compatibility with various processing engines.

Apache Iceberg stand in 2024 as the **decoupling of storage and compute** layers. Most storage systems now offer integration with cloud-based object storage solutions like S3, leveraging their inherent efficiency and elasticity. This approach allows data processing resources to scale independently from storage, leading to cost savings and enhanced scalability. Cockroachdb supporting S3 as storage backend, and **CockroachL** offering Long Term Kafka topic data retention on S3 further exemplifies this trend, highlighting the growing use of data lakes in cost-effective, long-term storage solutions.

One of the hottest developments in 2023 was the rise of **open table formats**. These frameworks essentially act as a table abstraction and virtual data management layer sitting atop your data lake storage and data layer as depicted in the following diagram.

Data Lakehouse abstraction layers

The open table format space is currently dominated by a fierce battle for supremacy between the following three major contenders:

- **Apache Hudi**: Initially developed and open sourced by **Uber**, with main design goal for near-real-time data updates and CDC ingestion.
- **Apache Iceberg**: Born from **Netflix's** engineering team.
- **Delta Lake**: Created and open sourced by **Databricks**, with seamless integration with the Databricks platform.

The funding received by the leading SaaS providers in this space in 2023 — **Databricks**, **Tabular**, and **Confluent** — emphasizes market interest and their potential to influence data management on data lakes.

Moreover, a new trend is now unfolding with the emergence of **unified data lakehouse layers**. **OneTable** (recently open sourced by Dremio) and **XTable** (currently non-open source offering from Onehouse) are the first two projects which were announced last year. These tools go beyond traditional table formats, aiming to provide a unified layer that can interoperate under a single umbrella. This approach promises to enhance a universal format while ensuring data in processing engines in their preferred formats, leading to increased flexibility and agility.

3. Data Integration

The data integration landscape in 2023 witnessed not only continued dominance from established players like **Apache NiFi**, **Airbyte**, and **Debezium**, but also the emergence of promising tools like **Apache Inlong** and **Apache SeaTunnel** offering compelling alternatives with their unique strengths.

Meanwhile, **Streaming CDC** (Change Data Capture) has further matured, fueled by active development in the Kafka ecosystem. **Kafka connect** and **Debezium** plugins have become go-to choices for near-real-time data capture from transactional databases. The ability to leverage existing infrastructure for deployment using Kafka as the primary stream processing engine.

Beyond traditional databases, tools like **CloudQuery** and **Steampipe** are simplifying data integration from APIs, providing powerful solutions for ingesting data from diverse sources, which further highlights the growing importance of flexible integration with cloud-based services.

As the scales of event and messaging infrastructures, **Apache Kafka** maintains its strong position, though challenges like **Redpanda** are closing the gap. Redpanda's 600 million Series C funding in 2023 shows the growing interest in alternative message brokers offering low latency and high throughput.

4. Data Processing & Computation

The world of Stream processing continued to lead in 2023 with **Apache Spark** and **Apache Flink** remain the reigning champions, however Apache Flink made some serious headways in 2023. Cloud giants like **AWS** and **Alibaba** jumping on board with Flink as a service offerings, and Confluent's acquisition of Immerok for its own fully managed Flink as a service offering, show the momentum behind this powerful engine.

In the Python ecosystem, data processing libraries such as **Ibis**, **Dask**, **polars**, and **Ray** are available for exploiting multi-core processors. These parallel execution libraries further unlock possibilities for analyzing massive datasets within the familiar Python environment.

5. Workflow Management & DataOps

The workflow orchestration landscape is actually quite vibrant with several active projects competing, filled with established heavyweights and exciting newcomers.

Veteran tools such as **Apache Airflow** and **Dagster** are still going strong and remain a widely used engine around the recent hot debates in the orchestration space revolving around the concepts of declarative vs imperative paradigms of workflow orchestration engines. On the other hand in the past two years, GitHub has witnessed the rise of several compelling contenders, capturing significant traction. **Kestra**, **Temporal**, **Mage**, and **Windmill** are all worth watching, each offering unique strengths. Whether focusing on serverless execution, ease of use for data-focused tasks, or extensive file-type support, these newcomers are rising to the challenging world of modern data pipelines.

6. Data Infrastructure & Monitoring

The recent **Confluent Latin America**, confirms **Grafana**, **Prometheus** and **ELK** stack continue to dominate the observability and monitoring landscape. **Grafana Labs** itself has been quite active, introducing new open source tools like **Loki** for log aggregation and **Mimir** for long-term Prometheus storage to further strengthen its platform.

One area where open source tools seem less prevalent is **cluster management and monitoring**. This likely stems from the cloud integration trend, reducing the need for managing large on-premise data platforms. While the **Apache Ambari** project, once popular for managing Hadoop clusters, reached its end-of-life in 2021, its successor, the **Cloudera Manager** in EPL+ remains a potential open source hope for its future. However, its long term data remains uncertain.

In the resource scheduling and workload deployment, **Kubernetes** seems to be the preferred resource scheduling specially on cloud-based platforms.

7. ML Platform

Machine Learning Platform has been one of the most active categories with unprecedented rise and interest in **Vector databases**, specialized systems optimized for the storage and retrieval of high-dimensional data. As highlighted by DB-Engines' 2023 report, vector databases emerged as the most popular database category in the past year.

MLflow tools also play an increasingly vital role in scaling ML projects effectively, ensuring smooth operations and ML pipelines efficient management. As the complexity and scale of ML deployments continue to grow, MLOps tools have become indispensable for streamlining development, deployment, and monitoring of ML models.

8. Metadata Management

In recent years, metadata management has taken center stage, propelled by the growing need to govern and improve management and access to data. However, the lack of comprehensive metadata management platforms prompted such giants like Netflix, Lyft, Airbnb, Twitter, LinkedIn and Paypal to build their own solutions.

Despite this challenge, dedicated Metadata Parallel Processing (MPP) engines such as **Elastic**, **Impala**, **Presto** and **Trino** remain prominent within large data platforms, especially for petabyte-scale data.

Beyond traditional MPP engines, **uniform execution engines** are another trend gaining traction. Engines such as **Apache Linkin**, **Nikanto** and **Calro** provide a query and computation middleware between upper applications and underlying engines.

Other Lists
Data Engineering
Open Source
Analytics
Data Platforms
Data Lake
Streaming
Lakehouse
Lakehouse
Lakehouse

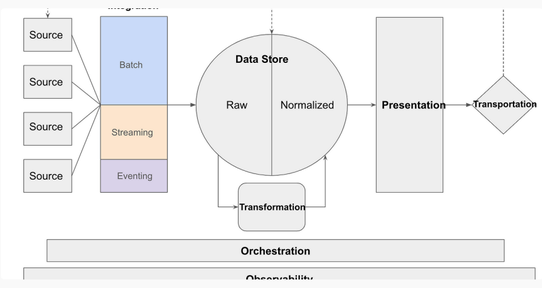
10 min read · Sep 19, 2023

8

6 min read · Oct 1, 2023

See all from Alireza Sadeghi

Recommended from Medium



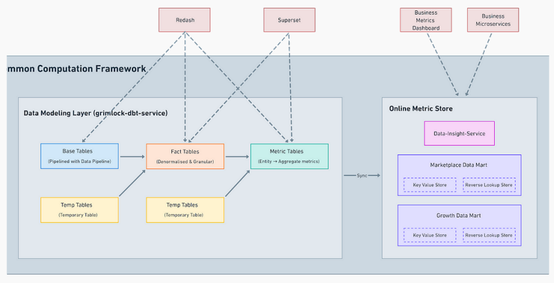
Dave Melillo in Towards Data Science

Building a Data Platform in 2024

How to build a modern, scalable data platform to power your analytics and data...

9 min read · 5 days ago

215 1



UC Blogger in Urban Company – Engineering

From Silos to Standardization: Leveraging DBT for a...

By—Shashank Chaudhary (Engineer, Platform)

8 min read · Feb 2

87

Lists

data science and AI
40 stories · 67 saves

Predictive Modeling w/ Python
20 stories · 893 saves

Icon Design
36 stories · 222 saves

Business
39 stories · 69 saves

Madison Schott in Dev Genius

3 Tips for Dimensional Data Modeling in dbt

dbt's built-in features make it easier than ever to produce high-quality data models

· 5 min read · 6 days ago

11

Mariusz Kujawski

Exploring the Medallion Architecture in Microsoft Fabric

The Medallion architecture stands out as one of the most popular frameworks for...

9 min read · 6 days ago

68 2

Yusuf Ganiyu

Top 5 Data Engineering Projects You Can't Afford to Miss

As we embark on the journey through 2024, the realm of data engineering is undergoing ...

7 min read · 4 days ago

122

Eylon Steiner in Snowflake

Snowflake New DataOps Related Features (2024)

How to manage Snowflake resources with Snowflake's Git Integration, EXECUTE...

7 min read · 5 days ago

32 1

See more recommendations