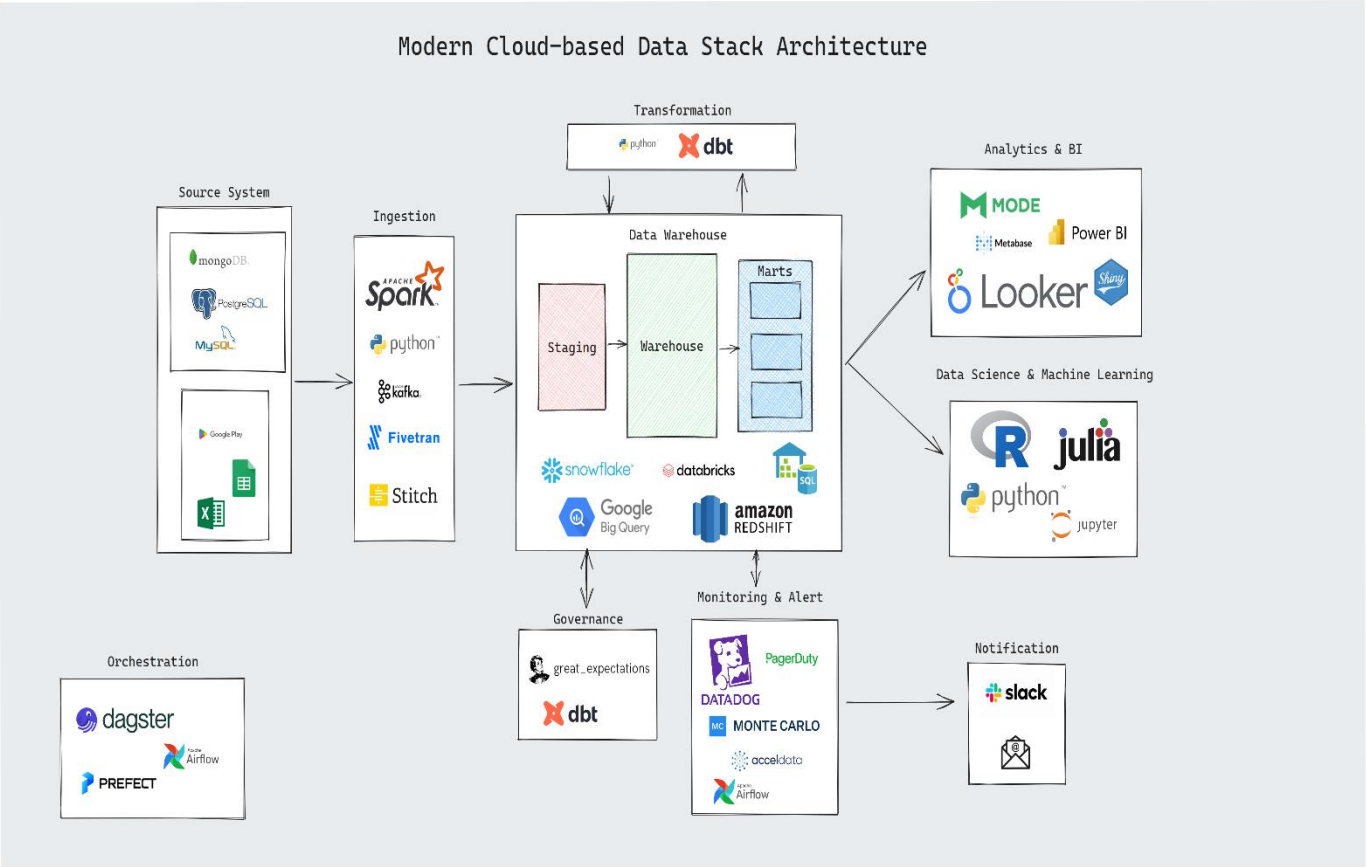
Modern Cloud-based Data Stack Architecture

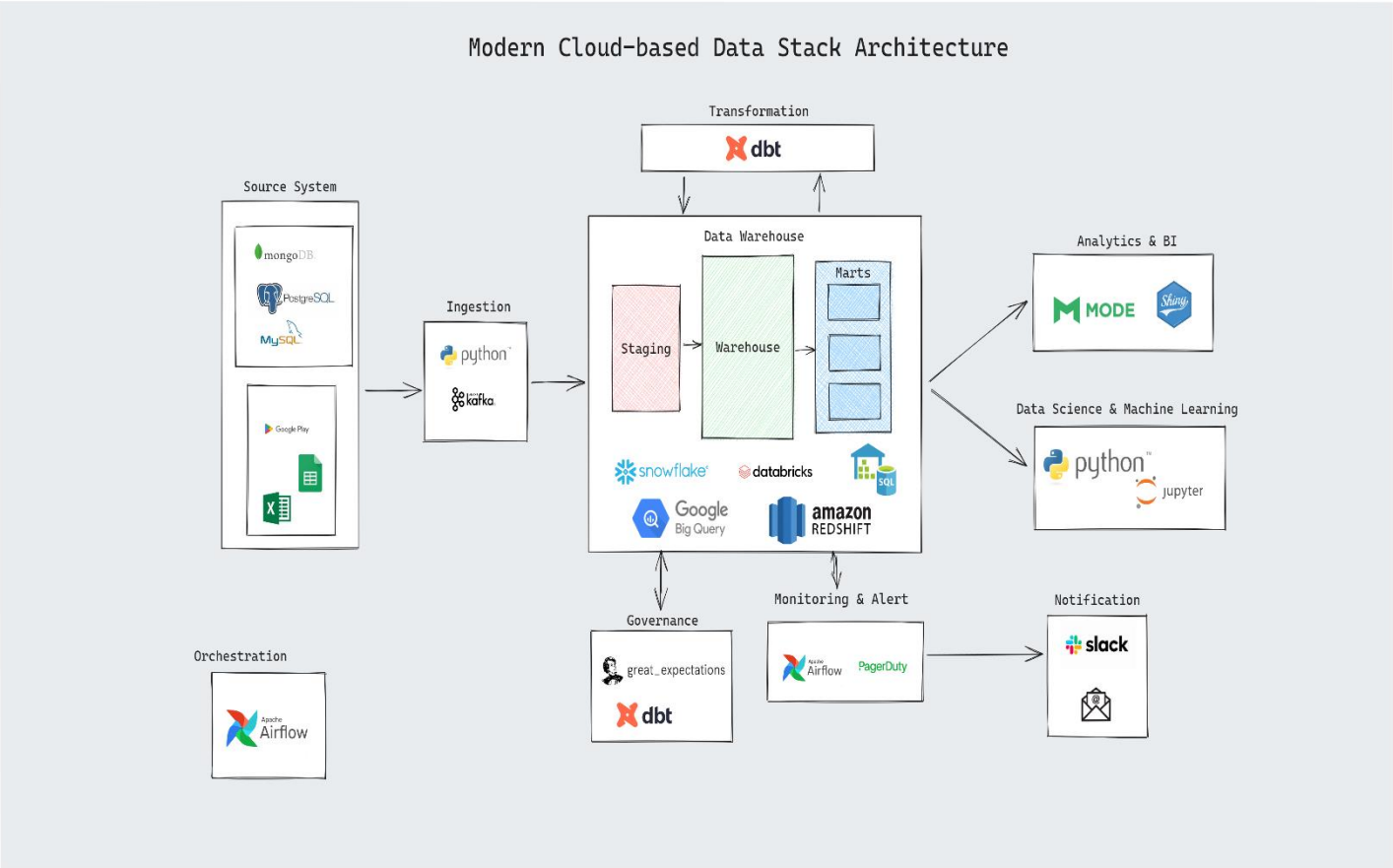


The Modern Data Stack Architecture

The modern data stack is a set of tools that together compromise a new approach to building data warehouse, from ingestion to transformation and serving of the data. They comprise of a set of independent tools and technologies that work together to enable businesses to collect, process, store and analyze data at scale.

These tools and technologies include the following:

- Tooling for ingesting data (e.g python, Stitch, fivetran, airbyte)
- A cloud data warehouse (e.g Snowflake, Redshift, BigQuery, Databricks, Azure Synapse)
- Transformation tools (e.g dbt, python)
- Orchestration tools like airflow, prefect, dagster
- Data visualization tools like Looker, Mode, Power BI, Shiny
- Data Quality and Governance like great expectations
- Monitoring and alert: datadog, pageduty, airflow, acceldata



Data Ingestion Layer:

- Snowflake: Snowflake provides scalability and elasticity, allowing you to scale resources based on your workload. Snowflake's pay-per-use pricing model helps optimize cost by only paying for the actual usage.
- Airflow: Apache airflow offers scalable and distributed task execution, allowing you to handle large volumes of data. Airflow's flexible scheduling capabilities enable efficient resource allocation and cost optimization.

Data Storage Layer:

- Snowflake: Snowflake's architecture is designed for scalability, providing automatic scaling of storage capacity based on your data growth. Snowflake's cloud-nativve storage and processing architecture optimize costs by minimizing storage and compute resources required.

Data Transformation Layer:

- dbt(Data Build Tool): dbt provides code-based, version-controlled data transformation capabilities. Its incremental processing and selective deployment allow for efficient updates, minimizing resource consumption and reducing costs. Dbt's collaborative features facilitate team collaboration and documentation.

Data Quality and Governance Layers:

- Great Expectations: Great Expectations helps ensure data quality and governance. By defining expectations and automated tests, you can validate data quality and detect issues early. Great Expectations' extensibility and integration with dbt enable collaborative data quality monitoring.

Analytics and Visualization Layer:

- Power BI: Power BI is a powerful data visualization and business intelligence platform. It allows users to create interactive dashboards, reports, and visualizations from various data sources.

Key takeaways:

Scalability:

- Snowflake: Snowflake's scalability allows for handling large datasets and concurrent workloads, ensuring performance and responsiveness. This supports scalability requirements for growing data volumes and user demands. As data volumes and workload increase, the cost of Snowflake usage may also increase. Proper monitoring and resource optimization are required to manage costs effectively.

Orchestration:

- Airflow: Airflow provides robust workflow orchestration allowing for complex data pipeline management, scheduling, and dependency handling. A managed service offers flexibility in task execution and resource allocation, supporting orchestration needs efficiently.

Collaboration:

- dbt's version-controlled approach enables collaborative development and documentation of data transformations.

Data Quality Monitoring:

- Great expectations allow users to define data expectations and validate data. With the combination of airflow and PagerDuty a monitoring and alerting system can be established.

Cost Optimization:

- The use of serverless and managed services helps optimize costs, as you pay only for the resources used and avoid infrastructure management overheads.