# Chose Effective Data Warehouse Technologies

Md. Obaidul Haque Sarker

Lead Data Architect

# Data Warehouse Evaluation Criteria

- **Pricing Model**: Pay-as-you-go, reserved instances, or consumption-based pricing.
- **Scalability**: Ability to scale compute and storage independently.
- **Performance**: Query execution speed and optimization.
- **Data Integration**: Support for ETL/ELT tools and real-time data ingestion.
- **Security & Compliance**: Encryption, access controls, and compliance certifications.
- **Ease of Use**: User-friendly interfaces and administration capabilities.
- **Machine Learning & AI Capabilities**: Built-in support for analytics and AI integration.

# Evolution of Data Warehouses Technologies

- **First Generation (1980s - Early 2000s) -** Traditional On-Premise Data Warehouses
- **Second Generation (Mid-2000s - 2010s) -** Big Data & Hadoop-Based Warehouses (2000s - 2010s)**.**
- **Third Generation (2010s - Present) -** Cloud Data Warehouses**.**
- **Forth Generation (Late 2010s – Present)** - Data Lakehouse & Real-Time Warehousing

# First Generation (1980s - Early 2000s) Traditional Data Warehouses

- **Architecture:** Monolithic, Enterprise Data Warehouse (EDW), On-premises.
- **Use Case:** Batch processing, structured data storage, and SQL-based analytics.
- **Key Technologies**: IBM Db2 Warehouse, Teradata, Oracle Exadata, MSSQL (OLAP) etc.
- **Characteristics :**
  - Built on relational databases (RDBMS) optimized for OLAP (Online Analytical Processing).
  - ETL-based batch processing to ingest structured data.
  - Star/Snowflake schema design to optimize queries.
  - Expensive proprietary hardware and software.
  - Slow, batch-oriented processing (daily or weekly updates).

# First Generation (1980s - Early 2000s) Traditional Data Warehouses (Cont.)

❑ **Pros:**

- ✔ High performance for structured data queries in low volume data.
- ✔ Mature and well-supported platforms.
- ✔ Strong ACID compliance and enterprise security.

❑ **Cons:**

- ✘ Expensive infrastructure and licensing costs.
- ✘ Poor scalability for unstructured and semi-structured data.
- ✘ Limited real-time analytics and distributed computing capabilities.
- ✘ Fixed Schema. Does not support big data processing.

# Second Generation (Mid-2000s - 2010s)
# Big Data & Hadoop

- **Architecture:** Distributed computing, batch processing (MapReduce), Data Lakes.
- **Use Case:** Handling large-scale unstructured data, data lakes, and batch processing.
- **Key Technologies:** Apache Hadoop, Apache Hive, Apache HBase, Cloudera Impala, Google Bigtable.
- **Characteristics:**
  - **MPP (Massively Parallel Processing)** architecture for high-speed analytics.
  - Distributed query execution across multiple nodes.
  - Optimized for structured **big data analytics**.
  - **Columnar storage** introduced for faster reads.
  - Still **hardware-intensive,** requiring dedicated appliances.

# Second Generation (Mid-2000s - 2010s)
# Big Data & Hadoop (Cont.)

❑ **Pros:**

☑ **Scalable and high-performance** for large datasets.

☑ **Faster query execution** due to columnar storage and MPP architecture.

☑ Optimized for **analytical workloads**.

❑ **Cons:**

✗ **Expensive and vendor-locked** solutions.

✗ **Complex infrastructure management**.

✗ Still batch-oriented, with **limited real-time capabilities**.

# Third Generation (2010s - Present)
# Cloud Data Warehouses

- **Architecture:** Serverless, MPP (Massively Parallel Processing), Real-time analytics.
- **Use Case:** Scalable, pay-as-you-go analytics, BI workloads, and ML integration.
- **Key Technologies:** Amazon Redshift, Google BigQuery, Snowflake, Microsoft Azure Synapse Analytics etc.
- **Characteristics:**
  - **Fully managed cloud services** (pay-as-you-go pricing).
  - **Separation of compute and storage** (Snowflake pioneered this).
  - Support for **real-time data ingestion and analytics**.
  - **Serverless query execution** (e.g., BigQuery).
  - **Integration with data lakes** and AI/ML tools.

# Third Generation (2010s - Present)
# Cloud Data Warehouses (Cont.)

❑ **Pros:**

☑ **Scalable and cost-efficient** with auto-scaling capabilities.

☑ **Faster deployment** (no hardware provisioning).

☑ **Supports semi-structured data (JSON, Avro, Parquet)**.

☑ **Real-time analytics** and interactive query execution.

❑ **Cons:**

✗ **Cloud dependency** (vendor lock-in risks).

✗ **Performance variability** due to shared cloud resources.

✗ **Data transfer costs** between storage and compute layers.

✗ Does not support Un-structured data.

# Forth Generation (Late 2010s - Present)
# Data Lakehouse & Real-Time Warehousing

- **Architecture:** Hybrid architecture combining Data Lakes and Data Warehouses.
- **Use Case:** Unifying structured, semi-structured, and unstructured data for AI/ML workloads.
- **Key Technologies:** Databricks (Delta Lake), Apache Iceberg, Apache Hudi, Trino (PrestoSQL), Dremio etc.
- **Characteristics:**

- **Hybrid approach** combining data lake storage with data warehouse querying capabilities.
- **Decoupled storage and compute** (open formats like Parquet, ORC).
- **ACID transactions** for consistency in big data environments.
- Optimized for **real-time and near-real-time analytics**.

Supports **unstructured, semi-structured, and structured data**.

# Forth Generation (Late 2010s - Present)
# Data Lakehouse & Real-Time Warehousing (Cont.)

❑ **Pros:**

☑ **Flexibility**: Supports both structured and unstructured data.

☑ **Cost-efficient**: Open-source options available.

☑ **Scalable & support Real-time analytics** and machine learning workloads.

☑ **Interoperability**: Works with multiple engines (Spark, Presto, Trino).

❑ **Cons:**

✗ **Complex setup and management** (requires expertise in multiple tools).

✗ Requires expertise to configure and optimize performance.

✗ Compatibility issues with legacy BI tools.

# Comparison of Data Warehouse Generations

| Feature | 1st Gen (Traditional DW) | 2nd Gen (Hadoop & Big Data) | 3rd Gen (Cloud DW) | 4th Gen (Lakehouse) |
|---|---|---|---|---|
| Data Type | Structured | Structured & Semi-structured | Structured & Semi-structured | All (Structured, Semi-structured, Unstructured) |
| Processing Model | Batch | Batch (MapReduce) | MPP (Massively Parallel Processing) | MPP + Streaming |
| Scalability | Limited | High | Elastic (Auto-scaling) | Infinite (Lakehouse) |
| Cost | High (CapEx) | Lower (Open-source) | Pay-as-you-go | Lower (Cloud & On-prem options) |
| Performance | High (for structured data) | Medium (batch latency) | High (real-time analytics) | Very High (AI & ML Ready) |
| ML & AI Support | Minimal | Limited | Moderate | Full (Lakehouse-optimized) |

# Cloud Based Data Warehouse (DW)

- AWS Redshift
- Google Cloud BigQuery
- Azure Fabric / Synapse
- Snowflake
- Databrick

# Feature Comparison of Cloud Based DW

| Feature | AWS Redshift | Google BigQuery | Azure Fabric | Snowflake | Databricks |
|---|---|---|---|---|---|
| Architecture | Shared-nothing, Columnar Storage | Serverless, Columnar Storage | Lakehouse architecture | Multi-cluster Shared Data | Lakehouse, Delta Lake Storage |
| Performance Optimization | Columnar compression, Spectrum | Auto-scaling, BI Engine | Optimized storage & compute | Automatic Clustering, Caching | Photon engine, Caching |
| Scalability | Manual cluster resizing | Fully managed, Auto-scaling | Auto-scaling compute & storage | Elastic scaling | Auto-scaling |
| Storage | S3-based, Redshift Spectrum | Columnar storage in Google Cloud | OneLake storage | Cloud-agnostic | Delta Lake storage |
| Storage Model | Columnar | Columnar | Lakehouse | Columnar | Lakehouse |
| Compute | Separated from storage | Serverless | Separated compute & storage | Separated compute & storage | Apache Spark-based |
| Pricing Model | Per node-hour | Pay-per-query | Pay-as-you-go, Reserved | Pay-per-compute & storage | Pay-per-use, Reserved pricing |
| Security | VPC, IAM, Encryption | IAM, Data Loss Prevention | Azure Active Directory, Encryption | Role-based, End-to-end encryption | Fine-grained access control |
| BI & Analytics Integration | Amazon QuickSight, Tableau | Looker, Tableau | Power BI, Synapse | Power BI, Tableau, Looker | MLflow, SQL Analytics |
| Machine Learning Support | SageMaker integration | BigQuery ML | Azure ML | Snowpark, Python, Scala | Databricks ML, AutoML |

# Feature Comparison of Cloud Based DW (Cont.)

| Feature | AWS Redshift | Google BigQuery | Azure Fabric | Snowflake | Databricks |
|---|---|---|---|---|---|
| Ease of Use | Moderate | Easy | Easy | Very easy | Moderate |
| Multi-Cloud Support | AWS only | Multi-cloud | Azure only | Multi-cloud | Multi-cloud |
| Data Sharing | Redshift Data Sharing | Data Transfer Service | Integrated | Secure Data Sharing | Delta Sharing |
| ETL & Data Ingestion | AWS Glue, Lambda | Dataflow, Pub/Sub | ADF, Synapse pipelines | Snowpipe for streaming data | Databricks Auto Loader |
| Performance | Columnar storage, optimized for SQL workloads | Fast query execution using Dremel | High performance for analytical queries | Micro-partitioning for efficiency | Optimized for AI/ML workloads |
| Best For | Traditional BI | Ad-hoc analytics | Enterprise | General BI & analytics | ML & AI Workloads |
| Pricing Model | Pay per instance & usage | Pay per query (on-demand) | Consumption-based pricing | Pay-as-you-go & reserved instances | Pay-per-use & reserved instances |

# ETL/ELT and Streaming Tools

- ETL / ELT: Airflow, Airbyte, DBT
- Streaming Pipeline: Apache Kafka, Apache Spark,  AWS Glue

# Pricing Comparison of Cloud Based DW

| Data Warehouse | Pricing Model | Storage Cost | Compute Cost | Free Tier Availability |
|---|---|---|---|---|
| AWS Redshift | Pay-per-node or Serverless | $0.024/GB per month | $0.25 per RPU | Yes (Redshift Spectrum) |
| Google BigQuery | Pay-per-query (serverless) | $0.02 per GB | $5 per TB | Yes (First 1TB free) |
| Azure Fabric | Pay-as-you-go, Reserved | $0.023 per GB | Based on DWU | Limited Free Tier |
| Snowflake | Pay-per-second usage | $23 per TB/month | $2-$6 per credit | Yes (trial credits) |
| Databricks | Pay-per-use, Reserved Pricing | $0.02 per GB | $0.15 per DBU | Yes (Community Edition) |

# Summary of Monthly Costs

(1TB Data Processed & Stored)

| Service | Storage Cost (1TB) | Compute (Query) Cost (1TB) | Total Estimated Cost |
|---|---|---|---|
| AWS Redshift | $24 | $791 | $815 |
| Google BigQuery | $20 | $6 | $26 |
| Azure Fabric | $23 | $5 | $28 |
| Snowflake | $23 | $20 | $43 |
| Databricks | $23 | $20 | $43 |

# Key Differentiators of Cloud Based DW

- **AWS Redshift:** Best suited for businesses already using AWS services, with strong integration and a balance between performance and cost.
- **Google BigQuery:** Ideal for serverless, pay-per-query workloads with built-in ML capabilities.
- **Azure Fabric:** Optimized for Microsoft ecosystem users, offering OneLake storage and seamless integration with Power BI.
- **Snowflake:** Strong multi-cloud capabilities, automatic scaling, and great for mixed workload analytics.
- **Databricks:** Best for data lakehouse architecture, real-time data processing, and AI/ML workloads.

# Recommendation for Cloud Based DW

- **For cost-sensitive users**: Google BigQuery (pay-per-query).
- **For Microsoft Azure users**: Azure Fabric integrates seamlessly with the Azure ecosystem.
- **For AI/ML-heavy workloads**: Databricks provides the best ML and big data capabilities.
- **For enterprises needing flexibility**: Snowflake offers strong multi-cloud compatibility.

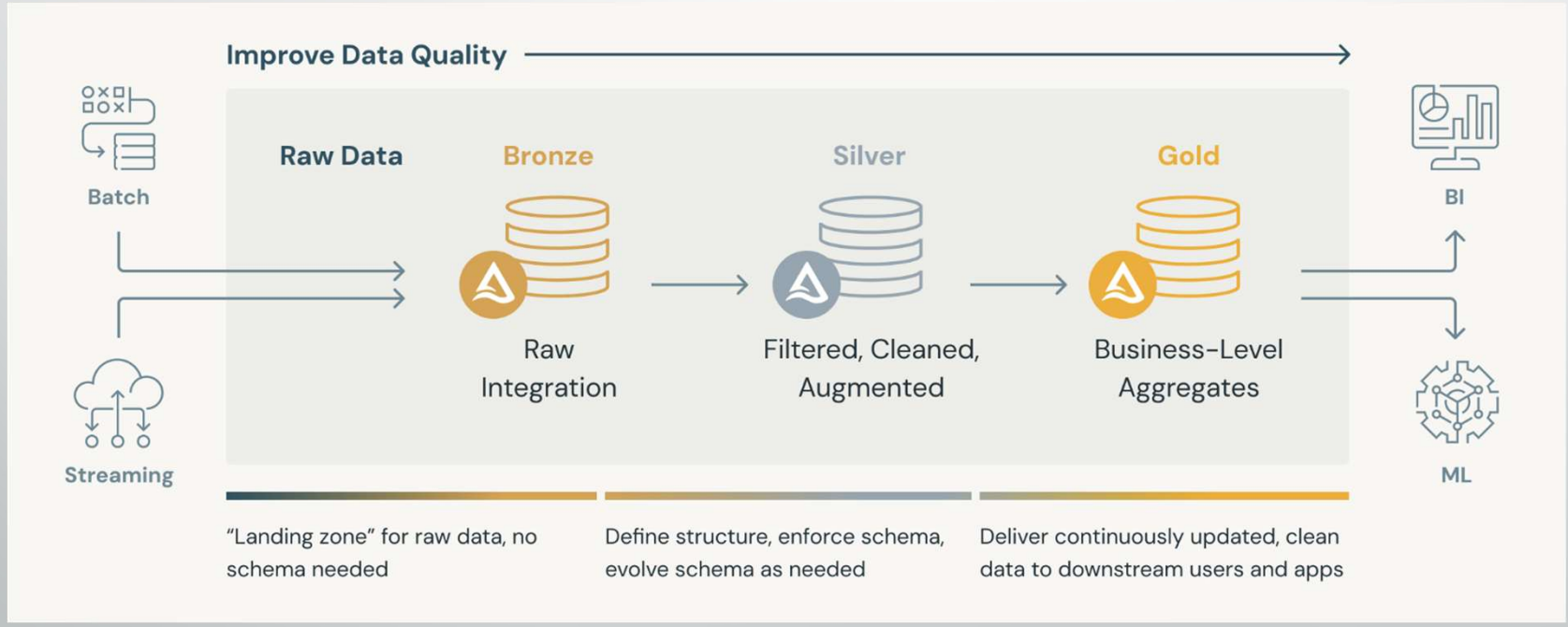
# On-Prem DW Technologies

| Category | Tools | Description |
|---|---|---|
| A. Data Lake | Apache Hadoop HDFS | Distributed storage system for big data. |
| | MinIO | S3-compatible object storage for on-premises cloud. |
| B. Data Lakehouse | Apache Iceberg | Table format for large-scale analytics in data lakes. |
| | Delta Lake | Lakehouse technology that brings ACID transactions to data lakes. |
| | Trino (PrestoSQL) | Query engine for data lakes and warehouses. |
| | Apache Hudi | Manages large-scale data updates on data lakes with transactional support. |
| C. Data Ingestion & ETL | Apache NiFi | Data flow automation for moving large datasets. |
| C.1 Batch Processing | Airebyte | Data Ingestion Tool |
| | Apache Airflow | Workflow automation tool for data pipelines. |
| C.2 Streaming Processing | Apache Kafka | Message broker for real-time data streaming. |
| | Apache Spark | Real-time analytics using Spark. |
| D. Data Governance | Apache Ranger | Security policy enforcement for Hadoop-based systems. |
| | Apache Atlas | Metadata and governance for data lakes. |
| E. Business Intelligence (BI) | Apache Superset | Advanced BI and data exploration platform. |
| | Metabase | Lightweight open-source BI tool for dashboards. |

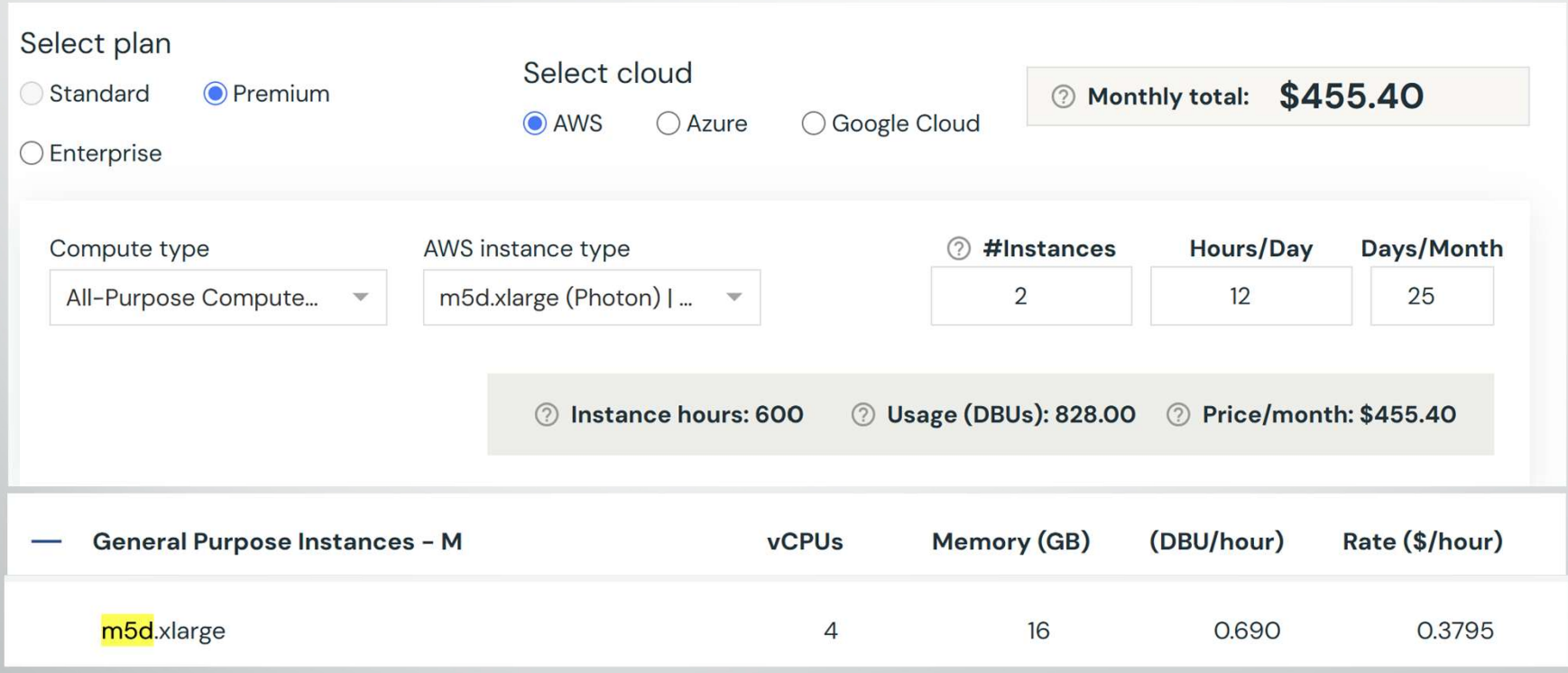
# Final Thoughts

- ✓ If we **need a traditional data warehouse**, **Cloud DWs (BigQuery, Snowflake, Redshift)** are the best choice for performance and ease of use.
- ✓ If we **require flexibility, open-source, and real-time analytics**, **Lakehouse architectures (Databricks, Apache Iceberg, Hudi)** are the future.
- ✓ If **on-premises** is a strict requirement, **Apache Iceberg, Apache Hudi, or Apache Hive** are viable solutions.

# Medallion Architecture



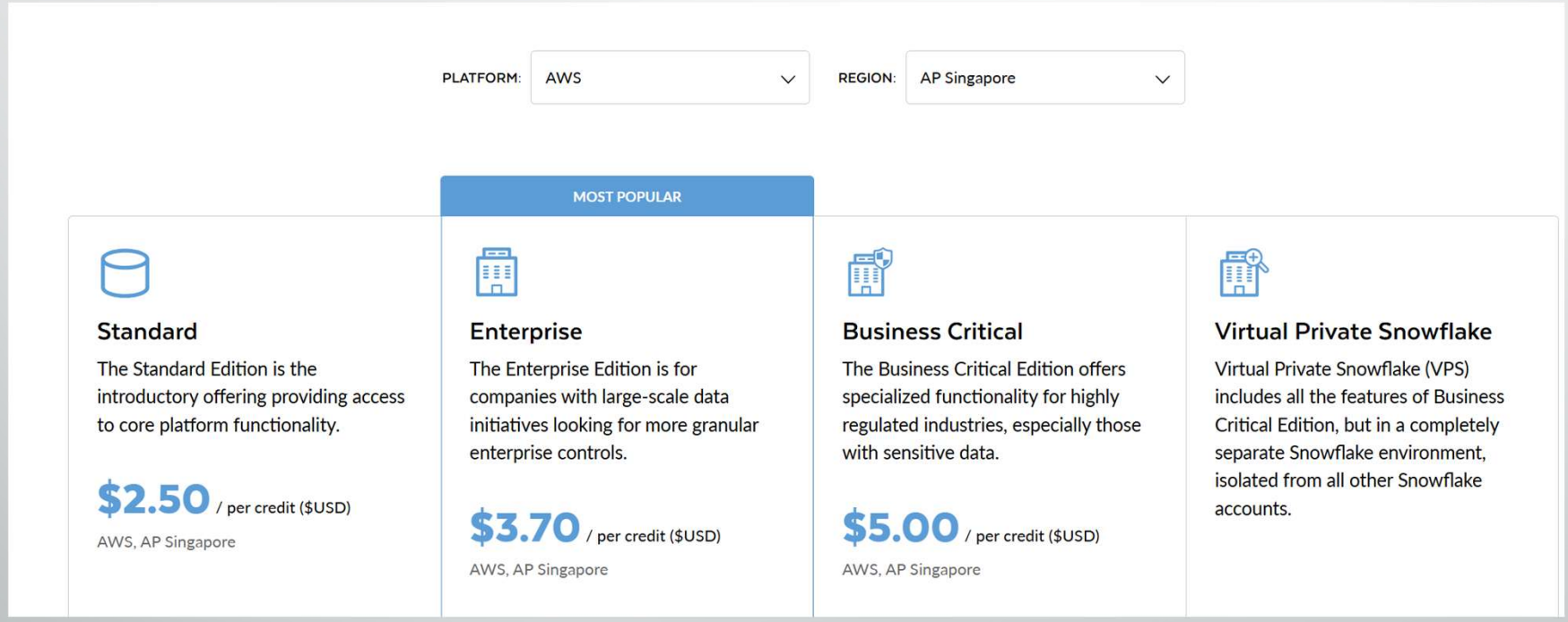
# Databrick Pricing




| General Purpose Instances - M | vCPUs | Memory (GB) | (DBU/hour) | Rate ($/hour) |
|---|---|---|---|---|
| m5d.xlarge | 4 | 16 | 0.690 | 0.3795 |

https://www.databricks.com/product/pricing/product-pricing/instance-types

# Snowflake Pricing Plan



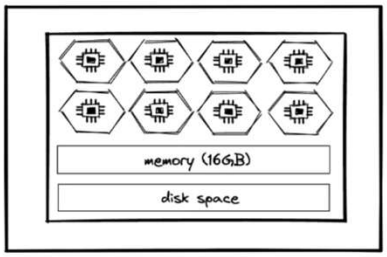

https://www.snowflake.com/en/pricing-options/

# Snowflake Pricing


https://snowflakecostcalculator.com/

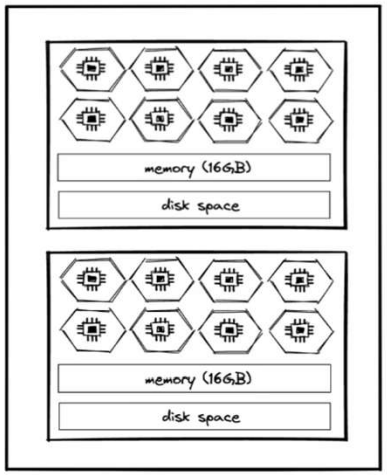
# Snowflake Virtual Warehouse Sizing

select.dev



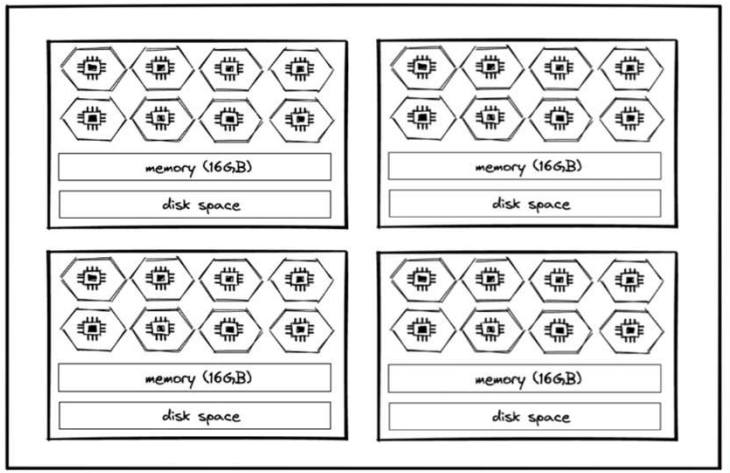

X-Small

- 1 node
- 8 cores/threads
- 16GB RAM
- 200GB local disk




Small

- 2 nodes
- 16 cores/threads
- 32GB RAM
- 400GB local disk




Medium

- 4 nodes
- 32 cores/threads
- 64GB RAM
- 800GB local disk

# ETL/ELT/Streaming Tools Pricing

AWS – Singapore Region

| Resource | Machine Type | Config | Price/Month (USD) | Purpose |
|---|---|---|---|---|
| EC2 Machine | t4g.xlarge | 4C, 16G, 100G | 100 | Airflow Server |
| AWS EMR | C4.4xlarge, 3 Nodes (1 master, 2 data) | 4C, 16G, 100G | 459.9 | Hadoop, Kafka, Spark |
| AWS EC2 | t4g.xlarge | 4C, 16G, 100G | 100 | Airbyte Server |
| AWS S3 | | 1 TB | 25 | Storage |
| | | **Total Price** | **684.9** | |

# Estimated Total Price

| Data Warehouse | DW Price/Month | ELT/ETL/Streaming Price/Month | Total Price |
|---|---|---|---|
| Databrick | $455 | $685 | **$1140** |
| Snowflake | $2000 | $685 | **$2685** |
| | | | |

# Engage in Questions & Discussion

### Open the Floor

Invite attendees to ask questions and seek clarifications regarding the presentation.

### Encourage Discussion

Foster conversations around specific needs and concerns related to data warehousing and analytics solutions.

### Gather Insights

Collect feedback from the audience to aid in enhancing decision-making processes.