Database Comparison – Sprint 2

Riley Shannon – 220432885

Table of Contents

Introduction

When choosing a database, it is essential that thorough research is done on the database selection and to have a list in mind. In sprint one I (Riley Shannon) constructed a document outlining a list of options of databases and what they are like. From this list and a few more we were able to come down to a small list to do a deep dive on and research their capabilities. These are:

- MongoDB
- Airbyte
- PostgreSQL
- PySpark

From these, below are their listed advantages, disadvantages, what they are best at and our most optimal solution.

MongoDB

MongoDB is a database system that is best at handling unstructured data out of the 4. This means that it is ideal for use cases involving large amounts of unstructured or semi-structured data, such as content management systems, IoT applications, and real-time analytics.

Advantages:

- **Schema Flexibility:** MongoDB is a NoSQL database, which means it handles unstructured data very well. It stores data in flexible, JSON-like documents, allowing for dynamic schemas.
- **Scalability:** Designed for horizontal scaling, MongoDB can handle large volumes of data across distributed clusters.
- **Ease of Use:** Its document model aligns well with modern, object-oriented programming, making it intuitive for developers.
- **Rich Query Language:** MongoDB provides a powerful query language, allowing complex queries and aggregations.

Disadvantages:

- **Consistency Trade-offs:** As a NoSQL database, MongoDB may face challenges with strong consistency, especially in distributed setups.
- **Complex Transactions:** While it supports multi-document ACID transactions, it's not as mature or straightforward as in traditional relational databases.
- **Memory Usage:** MongoDB can be memory-intensive, requiring careful resource management.

Airbyte

Airbyte out of all of these is the best for organisations looking to centralize data from multiple sources into a single database or data warehouse.

Advantages:

- **Data Integration:** Airbyte is an open-source ETL (Extract, Transform, Load) tool that excels at connecting various data sources and destinations.

- **Flexibility:** Supports a wide range of connectors, making it easy to integrate with many databases and data warehouses.
- **Customizable:** Highly customizable and extendable, allowing users to build custom connectors.
- **Open-Source:** Free to use with a strong community, though enterprise features are available.

Disadvantages:

- **Not a Database:** Airbyte is not a database but rather a data integration tool, meaning it cannot store or query data on its own.
- **Setup Complexity:** Initial setup and configuration can be complex, especially for non-technical users.
- **Resource Intensive:** Can be resource-intensive depending on the volume of data being processed.

PostgreSQL

PostgreSQL is best at relation data management where the applications require strong ACID compliance, complex queries and transactional integrity, for example financial systems, ERP and CRM

Advantages:

- **Relational Database Excellence:** PostgreSQL is a powerful, open-source relational database management system (RDBMS) known for its reliability and robustness.
- **ACID Compliance:** Offers full ACID (Atomicity, Consistency, Isolation, Durability) compliance, ensuring reliable transactions.
- **Extensibility:** Supports custom functions, data types, and even procedural languages, making it highly customizable.
- **Advanced Features:** Includes advanced indexing, full-text search, and JSON support for semi-structured data.

Disadvantages:

- **Complexity:** The richness of features can be overwhelming for new users, and complex queries can sometimes require significant tuning.
- **Vertical Scaling:** While PostgreSQL can scale vertically well, it has limitations when it comes to horizontal scaling compared to NoSQL databases like MongoDB.
- **Performance:** For massive datasets, performance can degrade unless optimized effectively.

PySpark

PySpark is best out of the 4 for big data analytics where there is large scale processing complex transformations, and machine learning pipelines in environments like data lakes.

Advantages:

- **Big Data Processing:** PySpark is a Python API for Apache Spark, making it excellent for processing large datasets in a distributed computing environment.
- **Scalability:** Handles massive amounts of data by distributing processing across a cluster of machines.
- **Integration with Big Data Tools:** Seamlessly integrates with Hadoop, Hive, and other big data tools, making it suitable for complex data pipelines.
- **Machine Learning:** Comes with built-in support for machine learning through Spark MLlib.

Disadvantages:

- **Complexity:** Setting up and managing a Spark cluster can be complex, and PySpark requires a good understanding of both Python and distributed computing.
- **Not a Database:** Like Airbyte, PySpark is not a database but a data processing engine, meaning it requires integration with other tools for storage and retrieval.
- **Latency:** PySpark is designed for batch processing, so real-time processing may introduce latency issues.

Conclusion:

If you need a **general-purpose relational database** with strong transactional support, advanced querying capabilities, and the ability to handle structured and semi-structured data, **PostgreSQL** is the best choice. It balances performance, features, and reliability, making it a versatile solution for a wide range of applications.

MongoDB is the better choice if your application primarily deals with unstructured data or requires high scalability and flexibility. It's particularly suited for modern web applications, real-time analytics, and other scenarios where data schemas may evolve rapidly.

Airbyte and **PySpark** are not databases but play critical roles in data integration and processing. Use **Airbyte** for integrating and moving data between systems, and **PySpark** for processing large datasets in a distributed environment, particularly when working with big data.

Conclusion

- **Best for Relational Data:** PostgreSQL
- **Best for Unstructured Data:** MongoDB
- **Best for Data Integration:** Airbyte
- **Best for Big Data Processing:** PySpark

Choosing the right tool depends on your specific use case, data structure, and performance requirements. If your primary need is for a robust, general-purpose database, **PostgreSQL is often the best overall choice.**