**DAT 260 Module Three Assignment  
Big Data Analysis Tools**

## Tool Comparison Table

| Tool | Strengths | Weaknesses | Best Used |
| --- | --- | --- | --- |
| Hive | * Data Processing – SQL – like Interface & Scalability * Project Management – Versatility, Visualization & centralized communication | * Project Management – Limited Mobile Functions & Task Dependencies * Data Processing – Limited Real- Time Processing & Performance | * Project Management * Big Data Processing * IT and Software Development -best industry * Analytics- Data Warehousing &   Ad-hoc querying |
| Spark | * Speed * Versatility * Machine Learning Capabilities * Ease Of Use | * Lack of Real Time Processing * Complexity * Manual Optimization | * Machine Learning * Data Processing * Exploratory Data Analysis * Tech, Finance & Healthcare – best industry |
| Flink | * Real – Time Processing * Ease of Use * Built- In Tolerance Systems | * Complexity * Memory Management * Lack of Unique Features | * Event Driven Applications * Data Analytics & Pipeline Applications * IT & Software Development & Data Analytics – best industry |
| Pentaho | * Real- Time Analysis & Visualizations * Open Source * Comprehensive Suite of product with focus on business Intelligence | * Limited Documentation * Slow updating evolution compared to other tools * Delayed support on through new versions | * Companies with 50-200 employees * Data Integration * Business Analytics * Dashboarding * IT & Software Development & Data Analytics – best industry |

## Reflection

As a data analytics professional, the selection of implements varies on the constraints of the job you are trying to accomplish and the industry in which you are employed. Reviewing the above tools the overall summary is :

Hive – is exceptional for the handling of large datasets and is beneficial for ad-hoc or as needed data analysis. Due to the SQL similar interface is makes it accessible(Sherrer, 2024).

Spark – is adaptable in that it can handle both batch processing and real-time data processing. It is formidable in machine learning and visualization making it a supreme choice for advance analysis (Wojciech-Gebis, 2022).

Flink – is a forerunner in real- time data processing and applications that are for at the moment result driven decisions. Optimal insights that require immediate attention (VnCallCenter, 2023).

Pentaho – focused on business intelligence it provides a comprehensive suite of products that cover all aspects of the data requirements from integration to dashboarding. Considered a one stop solution for business analytics (www.naukri.com, 2024).

In my new career industry, biomedical engineering and manufacturing, real- time processing and access to advanced analytics is critical. With the need to address changes on the spot, tools like Spark or Flink would be my choice. Being that they have machine learning capabilities this would be useful to predict failures, process optimization and quality control. However, due to the need for a strong reporting and dashboard services Pentaho would also be an option that requires consideration as it allows for a single point of development for data from various sources and permits for the visualization that is easily understood when presenting to stakeholders. Considering these factors, including the size of my organization, the volume and velocity of the data that is required to meet the demands of governmental regulations on both domestic and global standards, and the complexity of the manufacturing task it is imperative to look at the big picture. Analyzing current and future needs when deciding which tool or suite of tools is best for my industry is dependent on wise planning.

## Resources

Sherrer, K. (2024, March 21). Hive Review 2024: Pricing, Alternatives, Pros & Cons. TechRepublic. <https://www.techrepublic.com/article/hive-review/>

Wojciech-Gebis. (2022, November 17). What is Apache Spark? Architecture, Use Cases, and Benefits. Nexocode. <https://nexocode.com/blog/posts/what-is-apache-spark/>

Eliza Taylor. (2023, December 15). Advantages and Disadvantages of Apache Spark: A Complete Analysis. Www.theknowledgeacademy.com. <https://www.theknowledgeacademy.com/blog/advantages-and-disadvantages-of-apache-spark/>

VnCallCenter (2023, March 26). *The Pros and Cons of Using Apache Flink for Big Data Analytics*. Blog.VnCallCenter.com. <https://blog.vncallcenter.com/big-data-analytics/>

Kerner, S. M. (2019, April 25). Pentaho Business Analytics: Service Overview and Insight. Datamation. <https://www.datamation.com/big-data/pentaho-business-analytics-service-overview-and-insight/>

www.naukri.com. (2024). Code 360 by Coding Ninjas. Naukri.com. <https://www.naukri.com/code360/library/pentaho-and-big-data>

‌

‌