- Research 3 distributions that utilize the big data file systems approaches, and summarize the characteristics and provided functionality.

1. Apache Hadoop Distribution
    a. Components
        i. HDFS (Hadoop Distributed File System) – distributed storage of files across multiple machines, optimized for commodity hardware
        ii. YARN (Yet Another Resource Negotiator) – Hadoop's resource allocation and management mechanism for processing data in HDFS
    b. Features
        i. Fault Tolerance – Data is replicated across multiple nodes
        ii. High Availability – Replicated data allows for data availability even if one node fails
        iii. Simple Programming
        iv. Flexible Storage
        v. Low Cost
    c. Functionality
        i. Batch Processing – Hadoop uses MapReduce paradigm for large-scale data processing across nodes
        ii. Data Locality – Computation is conducted close to where the data is stored to minimize data transfer costs
        iii. Concurrent processing – YARN manages and schedules cluster resources to allow multiple data processing engines to run concurrently.
2. Google BigQuery
    a. Components
        i. Storage – structured table format
        ii. Ingestion – Capable of accepting data from multiple sources & data streams
        iii. Querying – standard SQL is used for querying and data analysis
    b. Features
        i. Serverless data warehouse
        ii. Fully managed – GCP handles all infrastructure including storage, scaling, and performance
        iii. Highly performant query execution – uses distributed architecture and massively parallel processing for fast SQL queries over petabytes of data
    c. Functionality

        i. SQL-Based Queries – utilizes google's ANSI SQL query language
        ii. Machine Learning – BigQuery allows users to build and train machine learning models directly within the platform using SQL
        iii. Data Analysis and Visualization – Offers integration with Google Data Studio and other BI tools for data visualization and insights
3. Cloudera Distribution
    a. Components
        i. Built on top of Hadoop and other tools
        ii. Includes cluster management, data governance and metadata management tools
    b. Features
        i. Offers fully managed service
        ii. Additional security and governance features
        iii. Professional customer support
    c. Functionality
        i. Offers same functionality as Hadoop + more managed tools and services.

- Research 3 distributions that utilize other NewSQL or NoSQL approaches, and summarize the characteristics and provided functionality.

1. Google Cloud Spanner (NewSQL)
    a. Features
        i. Distributed – Distributed globally across nodes
        ii. Fully Managed – All infrastructure is managed by Google Cloud Platform
        iii. Multi Dialect SQL Support – Offers Google's ANSI SQL and PostgreSQL
    b. Functionality
        i. Horizontal Scalability – Scalable across thousands of nodes
        ii. Global Replication – Supports synchronous replication across regions for high availability & low-latency read/writes
        iii. Strong Consistency – Offers strong consistency across distributed nodes
2. Cassandra (NoSQL)
    a. Features
        i. Distributed – distributed database across multiple servers
        ii. Eventual Consistency – distributed nodes are eventually consistent

iii. Wide-Column Store – Stored in column families allows for fast retrieval of large volumes of data

b. Functionality
   i. Horizontal Scalability
   ii. Fault Tolerance – replication offers fault tolerance and high availability
   iii. High Write Throughput – Optimized for heavy workloads

3. MongoDB (NoSQL)
   a. Features
      i. Schema-Free Data Model – stores data as flexible JSON-like documents
      ii. Horizontal Scalability – supports sharding which enables horizontal scaling
      iii. Eventual Consistency
   b. Functionality
      i. Document-Oriented Storage – allows for storage of hierarchical data, arrays, and nested documents
      ii. Indexing and Querying – Supports advanced indexing and querying to support rich queries, geospatial indexing and text searches
      iii. Replication and Sharding – Data replication ensures high availability and fault tolerance

- Compare and contrast how these technologies differ and the perceived benefits of each. Provide examples as necessary.

  All these options are horizontally scalable, whether they are fully managed or not. The MongoDB, Cassandra offer flexible schema-less data organization, whereas Spanner and BigQuery are strongly relational with schema. Hadoop and Cloudera are both distributed file systems, and BigQuery in the background utilizes a distributed file system. Spanner, MongoDB, and Cassandra, on the other hand, are distributed databases. All of these products are able to integrate with business intelligence tools.

GeeksForGeeks. (2023a, March 30). *Google Cloud Platform - Introduction to BigQuery*. GeeksforGeeks. https://www.geeksforgeeks.org/google-cloud-platform-introduction-to-bigquery/

GeeksForGeeks. (2023b, June 5). *Introduction to Hadoop*. GeeksforGeeks. https://www.geeksforgeeks.org/hadoop-an-introduction/

GeeksforGeeks. (2021, January 8). *Google Cloud Platform Introduction to Cloud Spanner*. GeeksforGeeks; GeeksforGeeks. https://www.geeksforgeeks.org/google-cloud-platform-introduction-to-cloud-spanner/

Introduction to Apache Cassandra. (2022, September 30). GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-apache-cassandra/

Saini, A. (2024, July 2). *What is MongoDB - Working and Features*. GeeksforGeeks. https://www.geeksforgeeks.org/what-is-mongodb-working-and-features/

Team Gyata. (2023, December 30). *Hadoop vs Cloudera: A Comprehensive Comparison | Big Data Processing*. Gyata.ai. https://www.gyata.ai/hadoop/hadoop-vs-cloudera

Tereshko, T., & Tigani, J. (2016, January 27). *BigQuery under the hood: Google's serverless cloud data warehouse*. Google Cloud Blog. https://cloud.google.com/blog/products/bigquery/bigquery-under-the-hood