Day12-9/25 Apache Cassandra, Apache HBase, Apache Hadoop

1. Apache Cassandra
   1. Free, open-source, distributed, wide column store, NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure
   2. Offers robert support for clusters spanning multiple data centers
   3. Asynchronous masterless replication allowing low latency operations for all clients
   4. Offers the distribution design of Amazon DynamoDB with the data model of Google’s Bigtable
2. Understanding what is a NoSQL Database
   1. NoSQL databases allow for unstructured data
   2. Offers a simple design, horizontal scaling, and extensive control over availability
   3. Does not require fixed schema
   4. Allows for easy replication
   5. Has a simple API
   6. However, they only support simply query language, i.e., are just ‘’eventually consistent’
   7. Does not support transactions
3. Benefits of using Apache Cassandra
   1. It offers highly-available service and **no single point of failure**
      1. True consistent access and availability
   2. It can handle massive volume of data
      1. Fast writes
   3. Its horizontal scalability
   4. **Flexible data storage** – can handle structured, semi-structured, and unstructured data, giving users flexibility with data storage
   5. **Flexible data distribution** – uses multiple data centers, which allows for easy data distribution wherever or whenever needed
   6. **Supports ACID** – The properties of ACID are supported by Cassandra
4. HBase
   1. An open source, non-relational, petabyte-scale, distributed database
   2. Based on Google’s BigTable that runs on top of Hadoop
   3. It provides for fast lookup of data because data is stored in-memory instead of on disk
   4. Integrates with Apache Hive
5. HBase has some advantage:
   1. Efficient storage of sparse data using column-based compression and storage
   2. Appropriate for high frequency counters (consistent reads & writes)
   3. High write & update throughput
   4. More integration with Hadoop
6. Apache Hadoop Ecosystem –has covered in Day2-HDFS
   1. MapReduce – Batch processing engine
   2. YARN – Resource management layer
   3. HDFS – Storage Layer

Resources cited:

https://en.m.wikipedia.org/wiki/Apache\_Cassandra

https://www.bmc.com/blogs/apache-cassandra-introduction