Development of Industries depends largely on their ability to find the errors that they have committed in the past and to take corrective measures in order to mitigate the losses as well as to increase the efficiency. Both these things can be achieved very easily when the data of the past can be used for the analysis and prediction purposes. But if we consider the data over a considerable amount of time, we find ourselves in the domain of big data. The industries that use big data storage, analytics and predictive modelling are not limited to but include e-commerce, scientific research, medical sciences, geology, weather survey, financial analysis and analysis of development schemes by the government. Hadoop uses map reduce architecture in order to distribute the computing between various clusters and provides flexibility for system scaling, also Hadoop being fault-tolerant and self-healing is an added advantage that it provides. Fault tolerance can be issued in four major ways – Batch Computing (Replication of data and rollback mechanism), Stream Computing (Passive & Active Standby, Upstream Backup), Spark and SDN Fault tolerance mechanisms.

Data Consistency is an important part of any big data system. The data consistency ensure that the latest changes are always provided to any process or thread that tries to use the data. But the CAP theorem states that there are three major components of any distributed data system – data consistency, system availability and tolerance to network partitions, and only two of the factors can be achieved at a time. There are two ways to observe consistency – client side and server side. The client-side consistency determines how the client sees the data updates while the server-side consistency determines the how the system updates the data and the amount of guarantee that the system can give about the data update. The major type of consistency that is used in the distributed data systems worldwide is the session consistency. The system guarantees both read and write consistency if the session is in progress. If the session expires of disconnects due to some issues, the consistency is no longer guaranteed. However monotonic read and monotonic write consistencies also can be used depending on the application and the trade off between the system availability and consistency. This is where the amazon dynamo comes in, the amazon dynamo is designed in such a way to allow the data to be spread over multiple places as well as the pick the optimum trade off between availability, performance, durability and consistency as per the application demands.

Data Lakes are used as a primary means for the big and distributed file systems. Data Lakes have the same major characteristics that any data pipeline should contain. The first step in any data lake is the ingestion. In this step, there is an ingestion framework that supports an array of input data types and structures. Data from various sources and in various forms is taken as an input to the system. The next major step is storage. Once the data is cleaned, it is stored in the relevant storage and its metadata is created. After cleaning the data, we perform data feature extraction and feature engineering on the data after which the new data is committed to the memory and the metadata is again saved. The Govern module of a data lake keeps the version control between the changes in the data in the form of meta data and is used when any changes must be made in an implemented and already running pipeline. The next major step is the analysis of the data using the machine learning algorithm which use the features that we have created from feature extraction and feature engineering steps. Finally, the last step of the data lakes is the Application part. The application part of the analysis that we have conducted can be a dashboard pointing out the trends or it could be smart applications (eg. User recommendations) or and end-to-end model that automates the process without any human involvement. The apply stage of the data lake produces some new data which can be given back to the system in order to increase the accuracy and reinforcing the learning strategy of the machine learning model.

References and Sources (By Paragraphs):

1. The Fault Tolerance of Big Data Systems | Xing Wu, Zhikang Du , Shuji Dai , and Yazhou Liu | School of Computer Engineering and Science, Shanghai University, Shanghai, China & Key Laboratory of Image and Video Understanding for Social Safety, Nanjing University of Science and Technology, Nanjing, China.
2. read-2009-Vogels-eventually-consistent.pdf| communications of the acm | january 2009 | vol. 52 | no. 1
3. Video – Data Lake