**INTRODUCTION:**

The concept of Big Data is used in many social networking websites like Facebook, Twitter etc., and many other streaming devices like Netflix, Amazon Prime and various online gaming devices. The data that can be in many forms, it can be unstructured hybrid type or in clean format. What is important is how this data can be stored, managed, and used in timely manner to extract information given its enormous and never-ending increasing size. New tools needed the for the organizations to manage data efficiently to fruit business benefits and to gain competitive edge in the products and services they provide.

**CASE STUDY: Microsoft Studios – XBOX**

**Challenges with Big Data:**

Today, majority of organizations are utilizing Big Data to extract bankable information from the data. Data is exponentially expanding, and to manage such enormous amount of data, Data warehouses and ETL systems require continuous expansion and improvements. Storage and processing solutions like Hadoop deployed physically on-sites face multiple issues. Hardware required to set it up require heavy and another challenge is to keep them going and ensure smooth operations which requires considerable amount of workforce which again has cost implications [1].

**Microsoft’s prevailing big data challenges on cloud systems.**

Microsoft is working on using machine level AI (machine learning) to manage Big Data problem by developing solutions that can automate management of the exponentially growing data and are cloud based. Cloud-computing systems use data storage differently than desktop machines. Cloud based solutions systems are still evolving models for storing and managing data. These include cloud-based databases like Microsoft’s SQL Azure, distributed file systems like Hadoop, and new data structures such as Azure’s queues and blobs. These cloud-based infrastructures have gained more popularity due to cost effectiveness and flexibility in big data handling. Microsoft Azure HDInsight is one such effective platform that provides opportunity to Microsoft to manage vast amounts of Xbox data [3]

Big Data on cloud systems still have lot on ongoing challenges. Some of these are: Complexities due to multiple components that need handling, testing, and validation. They need to be error free and require complex coding to perform smoothly. Also, Big Data require highly specialized languages and frameworks, The architecture can be complex and branched to many nodes, which require adaptable APIs that can be improved based on ongoing programming languages. Technology wise big data is still developing, like Hive and Pig are now deprecating, and newer technologies like spark is at its naïve stage. Managed services such as Azure Data Lake Analytics and Azure Data Factory are new emerging technologies. Security is one of the biggest challenges for big data pipeline. With large amount of data in a pool and being utilized by substantial number of applications creating and maintaining a secured gateway requires considerable amount of manpower and investments in creating and updating security packages. [2]

To address the above concerns, Microsoft has implemented **Azure HDInsight** to build big data management pipeline. The key features of HDInsight offers are given below [4]:

Text

Description automatically generated

**Features of HDInsight [4]**

**Source:** **https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview**

**AZURE HDInsight:**

By Microsoft definition, Azure HDInsight is defined as [3]:

“Azure HDInsight is a managed Apache Hadoop service that lets you run Apache Spark, Apache Hive, Apache Kafka, Apache HBase, and more in the cloud”.

Azure HDInsight has many utilities in [big data](https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview#what-is-big-data) processing. It can be used to process both historical and real-time data. These utilities are stated below: [4]:

* ETL: Extract Transform Load- Helps in ETL process which can be sent to data warehouse or analysis.
* Data Warehousing: provides platform to query unstructured data
* IOT (Internet of Things): Helps in capturing & processing live data.
* Data Science: provides platform to perform analysis on data
* Hybrid: Can be used to extend existing onsite big data bundle into cloud for analysis.

Graphical user interface

Description automatically generated with medium confidence

Data Warehousing capability- HDInsight [3]

**Source:** [**https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview**](https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview)

**How XBOX uses HDInsight:**

As stated in [5] Xbox uses HDInsight to manage its big data pipeline. This utilizes the general steps used for ETL process stated in [6]. The ETL process is done by uploading raw data into Azure Blob Storage and transforming data into indexed form or metadata store database using Hive. This is followed by creation of HDCluster and data is connected. Then a schema created to data in the data store. Then data is transformed and loaded.

Diagram

Description automatically generated with low confidence

ETL general -Apache Hadoop environment components support ETL at scale [6] https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-etl-at-scale

Graphical user interface, diagram

Description automatically generated

# Uses Apache Hive as an Extract, Transform, and Load (ETL) tool [7]

Source: https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-using-apache-hive-as-an-etl-tool

HDInsight clusters can support many programming languages, like Pig Latin, Hive QL and Spark SQL, Java, Python, Scala, etc. Any programming languages that is from open source community and used in big data processing can be operated on HDInsight..

**References:**

[1] https://www.motifworks.com/running-hadoop-big-data-workloads-using-azure-data-services/

## [2] <https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data>

[3] <https://docs.microsoft.com/en-us/azure/hdinsight/>

[4] <https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-overview>

[5] <https://azure.microsoft.com/en-us/blog/how-xbox-uses-hdinsight-to-drive-analytics-on-petabytes-of-telemetry-data/>

[6] <https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-etl-at-scale>

[7] <https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-using-apache-hive-as-an-etl-tool>