CMT 428 module 4

Module 4: Hadoop Ecosystem Tools

Tool	Description	Practical Example	Key Concepts
Hive	Data warehousing tool built on top of Hadoop for querying and managing large datasets.	**Example:** Analyze website traffic logs stored in HDFS to identify popular pages and user behavior patterns. Use HiveQL (similar to SQL) to query the data and generate reports.	HiveQL, Tables, Partitions, Buckets, External Tables, SerDe (Serializer/Deserializer)
Pig	High-level data flow language and execution framework for parallel computation.	**Example:** Process a large dataset of customer reviews to identify sentiment trends. Use Pig Latin scripts to load, transform, and analyze the data, and then store the results in HDFS.	Pig Latin, Data Flow, ETL (Extract, Transform, Load), UDFs (User Defined Functions)
HBase	NoSQL database that provides random, real-time read/write access to data stored in HDFS.	**Example:** Build a real-time recommendation system that suggests products to users based on their browsing history. Use HBase to store user profiles and product information, allowing for fast lookups and updates.	Column-oriented Storage, Key-Value Store, Regions, Column Families, Row Keys
Sqoop	Tool for efficiently transferring bulk data between Hadoop and relational databases.	**Example:** Import customer data from a MySQL database into HDFS for analysis. Use Sqoop to connect to the database, select the data to import, and transfer it to HDFS in a suitable format.	Connectors, Import/Export, Data Transfer, Incremental Loads
Flume	Distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.	**Example:** Collect and aggregate system logs from multiple servers into HDFS. Use Flume agents to capture the logs, process them, and store them in HDFS for further analysis.	Agents, Sources, Channels, Sinks, Interceptors
Oozie	Workflow scheduler system to manage Hadoop jobs.	**Example:** Create a workflow that involves importing data with Sqoop, processing it with Pig, and storing the results in Hive. Use Oozie to define the workflow,	Workflows, Actions, Control Flow Nodes, Action Nodes, Coordinators, Bundles

		schedule its execution, and monitor its progress.	

More examples

Hadoop Ecosystem Tool	**Description**	**Practical Examples**
Hive for Data Warehousing	Hive is a data warehousing tool built on top of Hadoop, allowing users to query and manage large datasets using a SQL-like language called HiveQL.	- **Example 1:** Use Hive to create a table: CREATE TABLE employee (id INT, name STRING, dept STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; - **Example 2:** Run a query to calculate the average salary from a table: SELECT dept, AVG(salary) FROM employee GROUP BY dept;
Pig for Data Analysis	Pig is a high-level platform for processing large datasets using a scripting language called Pig Latin, which simplifies complex data transformations.	- **Example 1:** Write a Pig script to load and filter data: data = LOAD 'input_data.txt' USING PigStorage(',') AS (id:int, name:chararray, age:int); filtered_data = FILTER data BY age > 30; - **Example 2:** Use Pig to perform a join operation between two datasets: joined_data = JOIN dataset1 BY id, dataset2 BY id;
HBase for NoSQL Data Storage	HBase is a distributed, scalable, and NoSQL database designed for low-latency, random read/write access to large datasets, providing real-time data storage.	- **Example 1:** Use HBase shell to create a table: create 'user_profiles', 'personal_info', 'activity' - **Example 2:** Insert data into an HBase table: put 'user_profiles', 'row1', 'personal_info:name', 'John Doe'
Sqoop for Data Import/Export	Sqoop is a tool used for efficiently transferring bulk data between Hadoop and relational databases (e.g., MySQL, Oracle) for import/export tasks.	- **Example 1:** Import data from MySQL into HDFS: sqoop import --connect jdbc:mysql://localhost/db --username user --password pass --table employees --target-dir /hdfs/employees_data - **Example 2:** Export data from HDFS to MySQL: sqoop export --connect jdbc:mysql://localhost/db --username user --password pass --table employees --export-dir /hdfs/employees_data
Flume for Data Ingestion	Flume is used for collecting, aggregating, and moving large	- **Example 1:** Set up a simple Flume agent to collect log data: Create a configuration

Hadoop Ecosystem Tool	Description	Practical Examples
	volumes of log data from various sources into Hadoop for storage and analysis.	file that specifies a source (log files), a channel (memory), and a sink (HDFS). - **Example 2:** Use Flume to ingest Twitter data into HDFS: Configure a Flume source to pull data from the Twitter API and store it in HDFS.
Oozie for Workflow Management	Oozie is a workflow scheduler that allows users to create directed acyclic graphs (DAGs) of workflows to automate the execution of Hadoop jobs.	- **Example 1:** Define an Oozie workflow that chains together a Hive job and a MapReduce job in an XML configuration file. - **Example 2:** Schedule a periodic Sqoop data import job using Oozie coordinator, which triggers the workflow every day.

Summary:

This table outlines the essential Hadoop ecosystem tools used for managing, processing, and analyzing big data. Each submodule is equipped with practical examples, illustrating how the tools are used in real-world scenarios, ranging from querying data in Hive to automating workflows with Oozie.

More

Hadoop Ecosystem Tool	Description	Practical Examples
Hive for Data Warehousing	Hive is a data warehousing tool built on top of Hadoop, allowing users to query and manage large datasets using a SQL-like language called HiveQL.	- **Example 1:** Use Hive to create a table: CREATE TABLE employee (id INT, name STRING, dept STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','; - **Example 2:** Run a query to calculate the average salary from a table: SELECT dept, AVG(salary) FROM employee GROUP BY dept; - **Example 3:** Partition data in Hive for efficient querying: CREATE TABLE sales_partitioned (id INT, date STRING, amount FLOAT) PARTITIONED BY (year INT) STORED AS TEXTFILE;
Pig for Data Analysis	Pig is a high-level platform for processing large datasets using a scripting language called Pig Latin, which	- **Example 1:** Write a Pig script to load and filter data: data = LOAD 'input_data.txt' USING PigStorage(',') AS (id:int, name:chararray, age:int); filtered_data =

Hadoop Ecosystem Tool	**Description**	**Practical Examples**
	simplifies complex data transformations.	FILTER data BY age > 30; - **Example 2:** Use Pig to perform a join operation between two datasets: joined_data = JOIN dataset1 BY id, dataset2 BY id; - **Example 3:** Perform a group and aggregation in Pig: grouped_data = GROUP data BY age; avg_age = FOREACH grouped_data GENERATE group, AVG(data.age);
HBase for NoSQL Data Storage	HBase is a distributed, scalable, and NoSQL database designed for low-latency, random read/write access to large datasets, providing real-time data storage.	- **Example 1:** Use HBase shell to create a table: create 'user_profiles', 'personal_info', 'activity' - **Example 2:** Insert data into an HBase table: put 'user_profiles', 'row1', 'personal_info:name', 'John Doe' - **Example 3:** Scan HBase table for a range of rows: scan 'user_profiles', {STARTROW => 'row1', STOPROW => 'row10'}
Sqoop for Data Import/Export	Sqoop is a tool used for efficiently transferring bulk data between Hadoop and relational databases (e.g., MySQL, Oracle) for import/export tasks.	- **Example 1:** Import data from MySQL into HDFS: sqoop import --connect jdbc:mysql://localhost/db --username user --password pass --table employees --target-dir /hdfs/employees_data - **Example 2:** Export data from HDFS to MySQL: sqoop export --connect jdbc:mysql://localhost/db --username user --password pass --table employees --export-dir /hdfs/employees_data - **Example 3:** Incremental import using Sqoop: sqoop import --connect jdbc:mysql://localhost/db --username user --password pass --table employees --incremental append --check-column id --last-value 100
Flume for Data Ingestion	Flume is used for collecting, aggregating, and moving large volumes of log data from various sources into Hadoop for storage and analysis.	- **Example 1:** Set up a simple Flume agent to collect log data: Create a configuration file that specifies a source (log files), a channel (memory), and a sink (HDFS). - **Example 2:** Use Flume to ingest Twitter data into HDFS: Configure a Flume source to pull data from the Twitter API and store it in

Hadoop Ecosystem Tool	**Description**	**Practical Examples**
		HDFS. - **Example 3:** Ingest Apache server logs with Flume: Configure a Flume source to monitor the Apache access log file and send the data to an HDFS sink for storage.
Oozie for Workflow Management	Oozie is a workflow scheduler that allows users to create directed acyclic graphs (DAGs) of workflows to automate the execution of Hadoop jobs.	- **Example 1:** Define an Oozie workflow that chains together a Hive job and a MapReduce job in an XML configuration file. - **Example 2:** Schedule a periodic Sqoop data import job using Oozie coordinator, which triggers the workflow every day. - **Example 3:** Use Oozie to orchestrate a workflow that involves a data ingestion step using Flume, followed by a Hive query for data aggregation and a MapReduce job for further processing.