Ramazan Abylkassov

Hive notes (continuously updated)

Task 1: Basic Hive Interaction Using Hue [Link to detailed MD file on GitHub]

Connected to Hue (localhost:8888, admin/admin). Created classicmodels database and external tables using \001 delimiter. Executed required queries: SELECT *, LIMIT 10, filtered/sorted query, GROUP BY aggregation, and exported results to local directory. Found HDFS location via DESCRIBE FORMATTED. Files are partially human-readable—plain text with non-printable \001 delimiter controlled by FIELDS TERMINATED BY property.

Task 2: Basic Hive Interaction Using Beeline [Link to detailed MD file on GitHub]

Connected via Beeline (beeline -u jdbc:hive2://localhost:10000 -n hive). Verified classicmodels exists, listed tables. Created newdb database and new_emp table using CTAS: CREATE TABLE new_emp AS SELECT * FROM classicmodels.employees. Verified with COUNT(*).

Task 3: Managed and External Tables Using Beeline [Link to detailed MD file on GitHub]

Inspected new_emp properties via DESCRIBE FORMATTED: Location (hdfs://namenode:8020/user/hive/warehouse/newdb.db/new_emp), File Type (TEXTFILE), Table Type (MANAGED_TABLE), numFiles (1). Files are human-readable (TEXTFILE format with \001 delimiter).

Dropped MANAGED table → HDFS directory deleted.
Explanation: Managed tables delete both metadata AND data because Hive "owns" the data.

Attempted CTAS with EXTERNAL → **Error:** CREATE-TABLE-AS-SELECT cannot create external table.
Explanation: External tables expect existing data; CTAS generates new data.

Workaround: Created as managed, then converted with ALTER TABLE new_emp SET TBLPROPERTIES ('EXTERNAL'='TRUE').

Dropped EXTERNAL table → HDFS files persist.
Explanation: External tables delete only metadata; data is preserved because Hive doesn't "own" it.

Manually removed with hdfs dfs -rm -r.

Task 4: Partitioned Tables [Link to detailed MD file on GitHub]

Retrieved DDL with SHOW CREATE TABLE customers. Created partitioned table with AVRO format (country column in PARTITIONED BY, not column list). Enabled dynamic partitioning and inserted 50 rows with partition column last in SELECT.

Q: Which EXPLAIN command shows partition details?
Answer: EXPLAIN EXTENDED

Q: Why is partitioning important for performance?
Answer: Partition pruning allows Hive to read only relevant partition directories instead of scanning the entire table, dramatically reducing I/O and processing time.

Q: What are the contents of the main directory?
Answer: Subdirectories for each partition value (format: partition_column=value)

Q: What are the subdirectory names?
Answer: country=USA, country=France, country=Spain, etc. Each contains AVRO data files.

Task 5: Hive ACID Tables [Link to detailed MD file on GitHub]

Enabled ACID settings (txn.manager, concurrency, bucketing, dynamic partition mode). Created transactional table with ORC format, bucketing, and transactional=true property.

Q: Which DESCRIBE operation checks DML support?
Answer: DESCRIBE FORMATTED — shows transactional: true flag indicating ACID support.

Performed DML operations:
- Inserted 3 rows
- Updated Adam's salary to 9000
- Inserted Alex
- Deleted John.

Final verification confirmed all changes: Sara (12000), Adam (9000), Alex (13000).