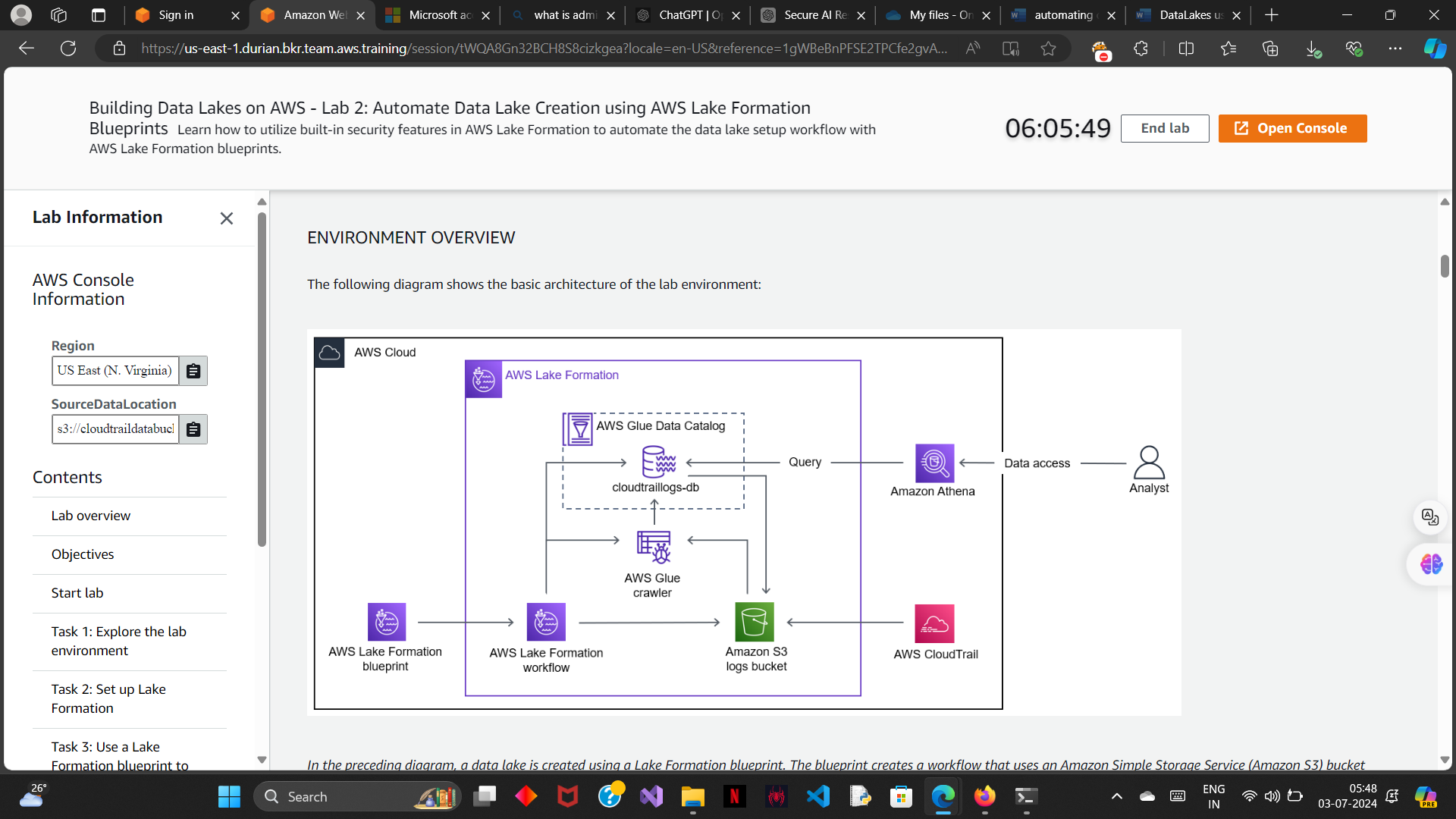
XXXXXXXXXXXXX

Labs Objectives -

2 line pitch -

AIM:

To use a workflow provided as an AWS Lake Formation blueprint to greatly simplify the creation of a data lake and ingestion of data. Lake Formation blueprints are workflows you can apply to an existing Lake Formation data lake. You can also apply them as a task in the setup and creation of a new data lake.



AWS CloudTrail is a service that provides a record of actions taken by a user, role, or an AWS service in Lake Formation. CloudTrail captures all Lake Formation API calls as events.

TASK 1 (lab env):

Important locations:

1> S3 --> cloud\_trail\_bucket --> data

2> Cloud Trail --> trails section --> lab cloud trail

TASK 2 (Set up the lake formation):

TASK 2.1 (registering s3 storage):

1> same as the previous lab

TASK 2.2 (create a database):

1> same as the prev lab.

TASK 3( use a lake formation blueprint to create an aws glue workflow) :

In this task, you use a Lake Formation blueprint to create an AWS Glue workflow that will automatically add new content to your data lake.

Workflows generate AWS Glue crawlers, jobs, and triggers to orchestrate the loading and updating of data. Lake Formation runs and tracks a workflow as a single entity. You can configure a workflow to run on-demand or on a schedule.

Workflows you create in Lake Formation are visible in the AWS Glue console as a directed acyclic graph (DAG). Each DAG node is a job, crawler, or trigger. To monitor progress and troubleshoot, you can track the status of each node in the workflow.

1> AWS lake formation --> ingestion section --> blueprints

2> follow the instructions:

* For **Blueprint type**, choose **AWS CloudTrail**.
* For **CloudTrail name**, choose **LabCloudTrail**.
* For **Start date**, choose today’s date.
* For **Target database**, choose **cloudtraillogs-db**.
* For **Target storage location**, add the **S3 bucket** from the left of instructions.
* For **Data format**, choose **Parquet**.
* For **Frequency**, choose **Run on demand**.
* For **Workflow name**, enter “lf-cloudtrail-workflow”
* .
* For **IAM role**, choose **LakeFormationWorkflowRole**.
* For **Table prefix**, enter “lab”

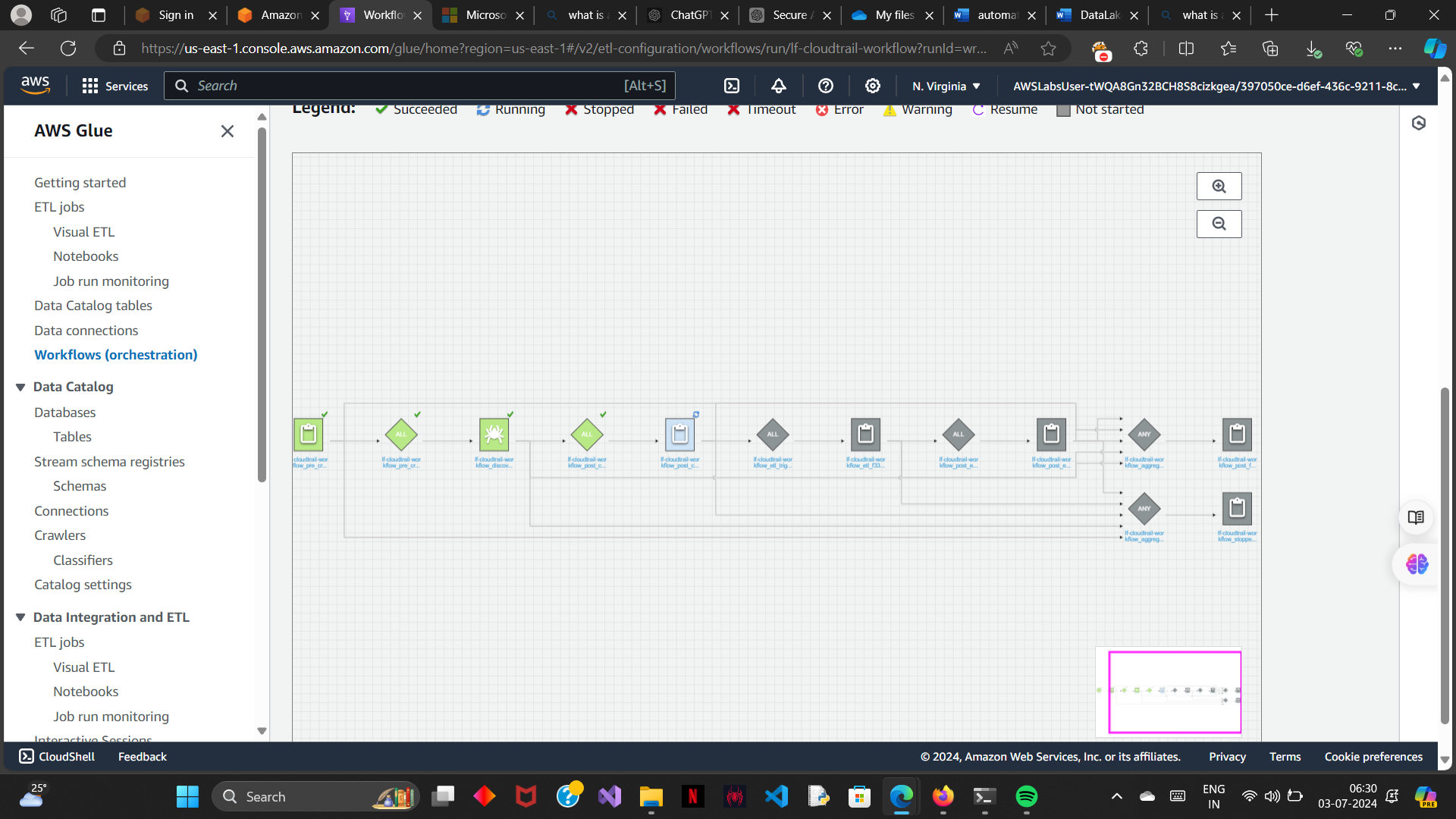
TASK 4 (run and monitor the workflow):

TASK 4.1(run the workflow):

1> click on the run button

2> to view the DAG:

AWS glue --> ETL section --> workflows --> if-cloudtrail-workflow --> history tab -->choose your instance --> view run details



TASK 4.2 (add custom workflow):

1> choose add workflow and provide a name.

2> In an AWS Glue workflow, you have four node types to choose from:

* **Start:** A node that starts the workflow.
* **Trigger:** A node that activates a job based on an event.
* **Job:** A node that completes work.
* **Crawler:** A node that discovers a source schema.

3> explaining the workflow:

* A **start** node that starts your workflow.
* A **pre\_crawler** job node that changes the workflow state to DISCOVERING.
* A **pre\_crawler** trigger node that starts the **crawler**.
* A **crawler** node that discovers the source schema.
* A **post\_crawler** trigger node that launches a **post\_crawler** job node.
* A **post\_crawler** job node that changes the workflow state to IMPORTING.
* An **etl** trigger node that launches an **etl** job node
* An **etl** job node that completes the ETL work.
* A **post\_etl** trigger that launches a **post\_etl** job node.
* A **post\_etl** job node that changes the workflow state to COMPLETED.

TASK 5 (validating the data lake setup results):

1> check out the details of the tables formed :

AWS lake formation --> data catalogue --> tables

Now here two tables are formed:

“\_lab\_cloudtrail” contains the cloud trail data before it was transformed into parquet format

“lab\_cloudtrail” contains the transformed data.

2> query the data lake results , cloud tail logs will be added as columns in the table

