#awscommunitybuilder bangladesh

Master Data Lakehouse with S3

Hands on Data Engineering Introduction Using AWS Services

Friday,
17 January 2025
8:00 PM - 9:00 PM BST

Online

With:

Partha Sarathi Kundu
Full-Stack Software Engineer
AWS Community Builder

Mim Ahmed Joy
DevOps Engineer
AWS Community Builder

AWS community builders

AWS COMMUNITY BUILDERS
PRESENTS:

DATA & CLOUD WORKSHOP

MASTER DATA LAKEHOUSE WITH S3

INTRODUCTION TO DATA LAKEHOUSE

- What is a Data Lakehouse?
- Why Choose a Lakehouse?
- How S3 is the best for Data Lakehouse?

AWS SERVICES OVERVIEW FOR LAKEHOUSE SOLUTION

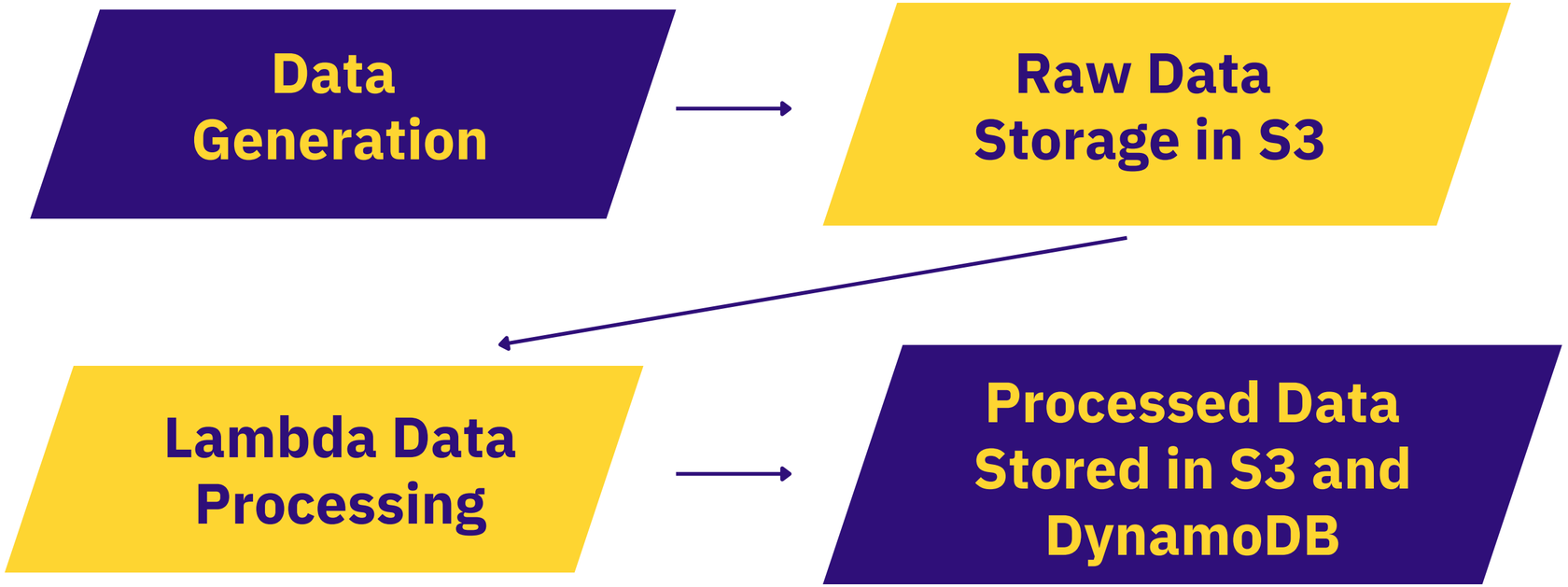
DATA INFRASTRUCTURE

- Amazon S3: Data lake storage.
- DynamoDB: NoSQL database for storing insights.
- Amazon Athena: Querying S3 data with SQL.
- Dash by Plotly: Interactive analytics dashboard.

CLOUD INFRASTRUCTURE

- AWS CDK: To Create AWS resources and build CI/CD pipeline
- AWS Services used: AWS S3, AWS DynamoDB, AWS Lambda, AWS IAM, AWS Athena, AWS CloudWatch, AWS Cloud Formation

DATA ARCHITECHTURE EXPLAINED



WHY DATA LAKEHOUSE?

SCALABILITY

- Leverages AWS services for automatic scaling.
- Serverless support storage and computation services.

COST-EFFECTIVENESS

- Pay-as-you-go pricing for S3, DynamoDB, Athena, and Lambda.
- Pay for only the Data transfer and storage amount used.

EASY TO INTEGRATE & CONFIGURE

- Flexible: Supports structured and unstructured data.
- Interactive Analytics: Real-time dashboards powered by DynamoDB and Dash.
- Configuration: Built in Trigger mechanism for multiple AWS Services.

HANDS-ON-DEMO

- Step 1: Generate Raw Data.
- Step 2: Configure CDK to build Cloud Infrastructure.
- Step 3: Upload Data to S3.
- Step 4: Configure S3 Trigger to Lambda
- Step 5: Monitor Lambda Processing via Cloudewatch logs.
- Step 6: Store Processed and formatted data in S3 & DynamoDB.
- Step 7: Query Data in Athena [Optional].
- Step 8: View Real-Time Dashboard.

PHASE 1 - SETUP

PROJECT INITIALIZATION

1. **Set Up the Project:**
   - **Clone the repository.**
   - **Create a Python virtual environment.**
   - **Install dependencies.**
2. **Generate Raw Data:**
   - **Use the provided script (data_generator.py) to simulate ecommerce events.**
3. **Organize Project Structure:**
   - **S3 folders: raw/, processed/.**

PHASE 2 - CONFIGURATION

AWS INFRASTRUCTURE SETUP

1. **Define Infrastructure with CDK:**
   - **S3 buckets, Lambda functions, DynamoDB tables.**
2. **Bootstrap and Deploy CDK Stacks:**
   - **Run cdk bootstrap.**
   - **Deploy using cdk deploy --all.**
3. **Verify Resources in AWS Console:**
   - **Check S3, DynamoDB, and Lambda.**

PHASE 3 - DATA

DATA LAKEHOUSE MANAGEMENT

1. **Create a Data LakeHouse with S3:**
   - **S3 will be created by CDK as ecommerce_data_lake;**
2. **Define a Trigger and destination:**
   - **S3 will trigger a Lambda to get the data and process it to store**
   - **Map the schema to processed data in S3.**
3. **Add Secoundary Data Storage:**
   - **DynamoDB is used to store processed Data and Structured Data for Visuals**
4. **Create Visuals:**
   - **Analyze user interactions and events for Visual**

PHASE 4 - DATA VISUAL

DATA DASHBOARD CREATION

1. **Set Up Dash App:**
    1. **Create a dashboard/ folder.**
    2. **Build interactive visualizations using Python [Dash].**
2. **Connect to DynamoDB:**
    1. **Fetch and aggregate insights dynamically.**
3. **Run the Dashboard Locally:**
    1. **Start the app: python app.py.**
    2. **View at http://127.0.0.1:8050/.**

Q&A SESSION

- Open the floor for participant questions.
- Encourage discussion around AWS services, use cases, and extensions.

GET CONNECTED WITH AWS CB & UG

AWS community builders

aws UG Bangladesh

THANK YOU

Hopefully this presentation can be useful for all of us. You can get conencted with me here on LinkedIn.

- www.linkedin.com/in/partha-sarathi-kundu/
- github.com/pxkundu/dataEngineeringTopics/tree/main/s3_lakehouse_workshop
- www.kundu.xyz