Introduction:

The increasing popularity of Over-The-Top (OTT) platforms like Netflix, Amazon Prime, Hulu, Disney Hotstar, etc., has led to a significant amount of data being generated related to user behavior, content consumption, and engagement. This data can provide valuable insights into user preferences, content performance, and much more. Building a data pipeline for OTT platform analysis can help organizations gain insights and make informed decisions. In this report, we will discuss building a data pipeline for OTT platform analysis using Netflix, Kaggle, IMDb, Rotten Tomatoes, Amazon Prime, Hulu, and Disney Hotstar, and storing the data in Amazon S3. We will also discuss how to trigger a Lambda function from a GitHub action, use AWS CloudFormation to create an EMR cluster that uses PySpark to clean the data and store it back in S3. Finally, we will discuss how Glue Crawler can be used to create table schemas, and Athena and Power BI can be used to gain insights from the data.

Building Data Pipeline:

1. Data Extraction: The first step in building a data pipeline for OTT platform analysis is to extract data from various sources like APIs, CSV files, databases, etc. For example, we can extract data from Netflix's APIs, Kaggle's CSV files, IMDb, Rotten Tomatoes, Amazon Prime, Hulu, and Disney Hotstar.
2. Storing Data in S3: Once the data is extracted, it can be stored in Amazon S3, which is a cost-effective and scalable solution for data storage.
3. Triggering Lambda Function from GitHub Action: We can use GitHub Actions to trigger a Lambda function that will be responsible for processing the data. The Lambda function can be triggered when a new file is uploaded to the S3 bucket.
4. Two Methods Used
   1. CloudFormation Template: The Lambda function can trigger a CloudFormation template that creates an EMR cluster using PySpark to clean the data and store it back in a new S3 bucket. PySpark is a distributed processing framework that can handle large datasets efficiently.
   2. AWS Glue: We can use AWS Glue, a fully managed extract, transform, and load (ETL) service, to clean the data and store it in a new S3 bucket. Glue can handle both structured and unstructured data and can transform the data using PySpark, a distributed processing framework.
5. Glue Crawler: We can use Glue Crawler to create table schemas in the data stored in S3. Glue Crawler is a fully managed ETL service that can automatically discover, catalog, and transform data.
6. Athena: Once the table schemas are created, we can use Athena to query the data stored in S3. Athena is an interactive query service that makes it easy to analyze data using standard SQL.
7. Power BI: Finally, we can use Power BI to create visualizations and gain insights from the data. Power BI is a business analytics service provided by Microsoft that can connect to various data sources, including Amazon S3.

Conclusion:

Building a data pipeline for OTT platform analysis can provide valuable insights into user preferences, content performance, and much more. By extracting data from various sources, storing it in Amazon S3, and using Lambda, CloudFormation, PySpark, Glue Crawler, Athena, and Power BI, we can build a scalable and cost-effective data pipeline that can help organizations make informed decisions.

