# Use cases summary

## NetApp Solutions

NetApp
October 20, 2023

# Table of Contents

# Use case summary

Previous: NetApp Spark solutions overview.

## Streaming data

Apache Spark can process streaming data, which is used for streaming extract, transform, and load (ETL) processes; data enrichment; triggering event detection; and complex session analysis:

- **Streaming ETL.** Data is continually cleaned and aggregated before it is pushed into datastores. Netflix uses Kafka and Spark streaming to build a real-time online movie recommendation and data monitoring solution that can process billions of events per day from different data sources. Traditional ETL for batch processing is treated differently, however. This data is read first, and then it is converted into a database format before being written to the database.
- **Data enrichment.** Spark streaming enriches the live data with static data to enable more real-time data analysis. For example, online advertisers can deliver personalized, targeted ads directed by information about customer behavior.
- **Trigger event detection.** Spark streaming allows you to detect and respond quickly to unusual behavior that could indicate potentially serious problems. For example, financial institutions use triggers to detect and stop fraud transactions, and hospitals use triggers to detect dangerous health changes detected in a patient's vital signs.
- **Complex session analysis.** Spark streaming collects events such as user activity after logging in to a website or application, which are then grouped and analyzed. For example, Netflix uses this functionality to provide real-time movie recommendations.

For more streaming data configuration, Confluent Kafka verification, and performance tests, see TR-4912: Best practice guidelines for Confluent Kafka tiered storage with NetApp.

## Machine learning

The Spark integrated framework helps you run repeated queries on datasets using the machine learning library (MLlib). MLlib is used in areas such as clustering, classification, and dimensionality reduction for some common big data functions such as predictive intelligence, customer segmentation for marketing purposes, and sentiment analysis. MLlib is used in network security to conduct real-time inspections of data packets for indications of malicious activity. It helps security providers learn about new threats and stay ahead of hackers while protecting their clients in real time.

## Deep learning

TensorFlow is a popular deep learning framework used across the industry. TensorFlow supports the distributed training on a CPU or GPU cluster. This distributed training allows users to run it on a large amount of data with lot of deep layers.

Until fair recently, if we wanted to use TensorFlow with Apache Spark, we needed to perform all necessary ETL for TensorFlow in PySpark and then write data to intermediate storage. That data would then be loaded onto the TensorFlow cluster for the actual training process. This workflow required the user to maintain two different clusters, one for ETL and one for distributed training of TensorFlow. Running and maintaining multiple clusters was typically tedious and time consuming.

DataFrames and RDD in earlier Spark versions were not well-suited for deep learning because random access was limited. In Spark 3.0 with project hydrogen, native support for the deep learning frameworks is added. This

approach allows non-MapReduce-based scheduling on the Spark cluster.

# Interactive analysis

Apache Spark is fast enough to perform exploratory queries without sampling with development languages other than Spark, including SQL, R, and Python. Spark uses visualization tools to process complex data and visualize it interactively. Spark with structured streaming performs interactive queries against live data in web analytics that enable you to run interactive queries against a web visitor's current session.

# Recommender system

Over the years, recommender systems have brought tremendous changes to our lives, as businesses and consumers have responded to dramatic changes in online shopping, online entertainment, and many other industries. Indeed, these systems are among the most evident success stories of AI in production. In many practical use cases, recommender systems are combined with conversational AI or chatbots interfaced with an NLP backend to obtain relevant information and produce useful inferences.

Today, many retailers are adopting newer business models like buying online and picking up in store, curbside pickup, self-checkout, scan-and-go, and more. These models have become prominent during the COVID-19 pandemic by making shopping safer and more convenient for consumers. AI is crucial for these growing digital trends, which are influenced by consumer behavior and vice versa. To meet the growing demands of consumers, to augment the customer experience, to improve operational efficiency, and to grow revenue, NetApp helps its enterprise customers and businesses use machine- learning and deep- learning algorithms to design faster and more accurate recommender systems.

There are several popular techniques used for providing recommendations, including collaborative filtering, content-based systems, the deep learning recommender model (DLRM), and hybrid techniques. Customers previously utilized PySpark to implement collaborative filtering for creating recommendation systems. Spark MLlib implements alternating least squares (ALS) for collaborative filtering, a very popular algorithm among enterprises before the rise of DLRM.

# Natural language processing

Conversational AI, made possible by natural language processing (NLP), is the branch of AI helping computers communicate with humans. NLP is prevalent in every industry vertical and many use cases, from smart assistants and chatbots to Google search and predictive text. According to a Gartner prediction, by 2022, 70% of people will be interacting with conversational AI platforms on a daily basis. For a high-quality conversation between a human and a machine, responses must be rapid, intelligent, and natural sounding.

Customers need a large amount of data to process and train their NLP and automatic speech recognition (ASR) models. They also need to move data across the edge, core, and cloud, and they need the power to perform inference in milliseconds to establish natural communication with humans. NetApp AI and Apache Spark is an ideal combination for compute, storage, data processing, model training, fine-tuning, and deployment.

Sentiment analysis is a field of study within NLP in which positive, negative, or neutral sentiments are extracted from text. Sentiment analysis has a variety of use cases, from determining support center employee performance in conversations with callers to providing appropriate automated chatbot responses. It has also been used to predict a firm's stock price based on the interactions between firm representatives and the audience at quarterly earnings calls. Furthermore, sentiment analysis can be used to determine a customer's view on the products, services, or support provided by the brand.

We used the Spark NLP library from John Snow Labs to load pretrained pipelines and Bidirectional Encoder

Representations from Transformers (BERT) models including financial news sentiment and FinBERT, performing tokenization, named entity recognition, model training, fitting and sentiment analysis at scale. Spark NLP is the only open-source NLP library in production that offers state-of-the-art transformers such as BERT, ALBERT, ELECTRA, XLNet, DistilBERT, RoBERTa, DeBERTa, XLM- RoBERTa, Longformer, ELMO, Universal Sentence Encoder, Google T5, MarianMT, and GPT2. The library works not only in Python and R, but also in the JVM ecosystem (Java, Scala, and Kotlin) at scale by extending Apache Spark natively.

Next: Major AI, ML, and DL use cases and architectures.

## Copyright information



## Trademark information