Top 5 Apache Spark Use Cases

Last Updated: 31 Mar 2022

Get 1250+ Data Science code snippets

GET NOW

PDF ⬇

To live on the competitive struggles in the big data marketplace, every fresh, open source technology whether it is Hadoop, Spark or Flink must find valuable use cases in the marketplace. Any new technology that emerges should brag some kind of a new approach that is better than its alternatives.

The creators of Apache Spark polled a survey on "Why companies should use in-memory computing framework like Apache Spark?" and the results of the survey are overwhelming –

- 91% use Apache Spark because of its performance gains.
- 77% use Apache Spark as it is easy to use.
- 71% use Apache Spark due to the ease of deployment.
- 64% use Apache Spark to leverage advanced analytics
- 52% use Apache Spark for real-time streaming.

Deploying auto-reply Twitter handle with Kafka, Spark and LSTM

Downloadable solution code | Explanatory videos | Tech Support

Start Project

Start Your First Project
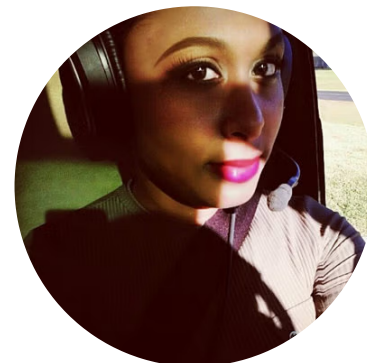### Learn By Doing

Email

Phone

Select Project ▼

START PROJECT

What Users are saying..

Camille St. Omer

Artificial Intelligence Researcher, Quora 'Most Viewed Writer in 'Data Mining'

in

I came to the platform with no experience and now I am knowledgeable in Machine Learning with Python. No easy thing I must say, the sessions are challenging and go to the depths. I looked at graduate... Read More

Relevant Projects

Machine Learning Projects

Data Science Projects

Python Projects for Data Science

Fast data processing capabilities and developer convenience have made Apache Spark a strong contender for big data computations. Apache Spark was the world record holder in 2014 "Daytona Gray" category for sorting 100TB of data. By sorting 100 TB of data on 207 machines in 23 minutes whilst Hadoop MapReduce took 72 minutes on 2100 machines. Fast data processing with spark has toppled apache Hadoop from its big data throne, providing developers with the Swiss army knife for real time analytics. Increasing speeds are critical in many business models and even a single minute delay can disrupt the model that depends on real-time analytics. In this blog, we will explore some of the most prominent apache spark use cases and some of the top companies using apache spark for adding business value to real time applications.

Table of Contents

You might also like

- Big Data Analytics Projects using Spark-Spark Projects

- Spark Use Cases in Advertising

-

"Only large companies, such as Google, have had the skills and resources to make the best use of big and fast data. There are many examples…where anybody can, for instance, crawl the Web or collect these public data sets, but only a few companies, such as Google, have come up with sophisticated algorithms to gain the most value out of it. Spark was designed to address this problem. Spark brings the top-end data analytics, the same performance level and sophistication that you get with these expensive systems, to commodity Hadoop cluster. It runs in the same cluster to let you do more with your data."- said Matei Zaharia, the creator of Spark and CTO of commercial Spark developer Databricks.

New Projects

Build Classification and Clustering Models with PySpark and MLlib

View Project

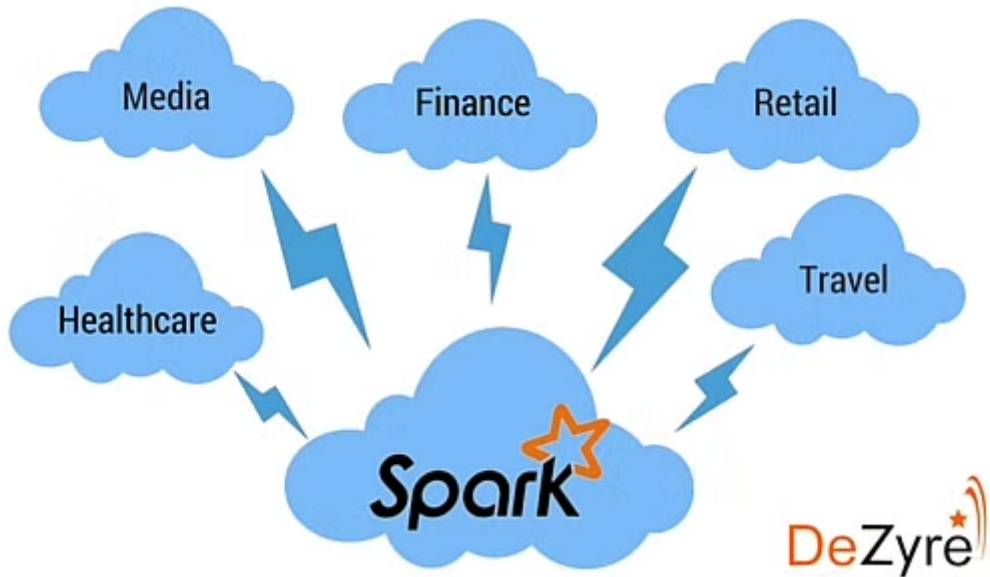**PySpark ETL Project-Build a Data Pipeline using S3 and MySQL**

View Project

PyTorch Pr⋯ Build a LST⋯ Classificatⁱ⋯

View P⋯

View all New Projects

Apache Spark Use Cases

Apache Spark is the new shiny big data bauble making fame and gaining mainstream presence amongst its customers. Startups to Fortune 500s are adopting Apache Spark to build, scale and innovate their big data applications. Here are some industry specific spark use cases that demonstrate its ability to build and run fast big data applications -

Spark Use Cases in Finance Industry

Right sidebar:

MNIST Dataset

Data Science Interview Questions

Python Data Science Interview Questions

Spark Interview Questions

Hadoop Interview Questions

Data Analyst Interview Questions

Machine Learning Interview Questions

AWS vs Azure

Hadoop Architecture

Spark Architecture

Tutorials

Data Science Tutorial

Snowflake Data Warehouse Tutorial for Beginners with Examples

Jupyter Notebook Tutorial - A Complete Beginners Guide

Best Python NumPy Tutorial for Beginners

Tableau Tutorial for Beginners - Step by Step Guide

MLOps Python Tutorial for Beginners -Get Started with MLOps

Alteryx Tutorial for Beginners to Master Alteryx in 2021

Free Microsoft Power BI Tutorial for Beginners with Examples

Theano Deep Learning Tutorial for Beginners

Banks are using the Hadoop alternative - Spark to access and analyse the social media profiles, call recordings, complaint logs, emails, forum discussions, etc. to gain insights that can help them make the right business decisions for credit risk assessment, targeted advertising and customer segmentation.

> **Get FREE Access to [Data Analytics Example Codes for Data Cleaning, Data Munging, and Data Visualization](#)**

Your credit card is swiped for $9000 and the receipt has been signed, but it was not you who swiped the credit card as your wallet was lost. This might be some kind of credit card fraud. Financial institutions are leveraging big data to find out when and where such frauds are happening so that they can stop them. They need to resolve any kind of fraudulent charges at the earliest by detecting frauds right from the first minor discrepancy. They already have models to detect fraudulent transactions and most of them are deployed in batch environment. With the use of Apache Spark on Hadoop, financial institutions can detect fraudulent transactions in real-time, based on previous fraud footprints. All the incoming transactions are validated against a database, if there a match then a trigger is sent to the call centre. The call centre personnel immediately checks with the credit card owner to validate the transaction before any fraud can happen.

Companies Using Spark in the Finance Industry

- One of the financial institutions that has retail banking and brokerage operations is using Apache Spark to reduce its customer churn by 25%. The financial institution has divided the platforms between retail, banking, trading and investment. However, the banks want a 360-degree view of the customer regardless of whether it is a company or an individual. To get the consolidated view of the customer, the bank uses Apache Spark as the unifying layer. Apache Spark helps the bank automate analytics with the use of machine learning, by accessing the data from each repository for the customers. The data is then correlated into a single customer file and is sent to the marketing department.
- Another financial institution is using Apache Spark on Hadoop to analyse the text inside the regulatory filling of their own reports and also their competitor reports. The firms use the analytic results to discover patterns around what is happening, the marketing around those and how strong their competition is.
- A multinational financial institution has implemented real time monitoring application that runs on Apache Spark and MongoDB NoSQL database. To provide supreme service across its online channels, the applications helps the bank continuously monitor their client's activity and identify if there are any potential issues.

Apache Spark ecosystem can be leveraged in the finance industry to achieve best in class results with risk based assessment, by collecting all the archived logs and combining with other external data sources (information about compromised accounts or any other data breaches).

compromised accounts of any other data breaches).

Spark Use Cases in e-commerce Industry

Information about real time transaction can be passed to streaming clustering algorithms like alternating least squares (collaborative filtering algorithm) or K-means clustering algorithm. The results can be combined with data from other sources like social media profiles, product reviews on forums, customer comments, etc. to enhance the recommendations to customers based on new trends.

Companies Using Spark in e-commerce

Industry

Shopify wanted to analyse the kinds of products its customers were selling to identify eligible stores with which it can tie up - for a business partnership. Its data warehousing platform could not address this problem as it always kept timing out while running data mining queries on millions of records. Shopify has processed 67 million records in minutes, using Apache Spark and has successfully created a list of stores for partnership.

Apache Spark at Alibaba

One of the world's largest e-commerce platforms Alibaba Taobao runs some of the largest Apache Spark jobs in the world in order to analyse hundreds of petabytes of data on its eCommerce platform. Some of the Spark jobs that perform feature extraction on image data, run for several weeks. Millions of merchants and users interact with Alibaba Taobao's ecommerce platform. Each of these interaction is represented as a complicated large graph and apache spark is used for fast processing of sophisticated machine learning on this data.

Apache Spark at eBay

eBay uses Apache Spark to provide targeted offers, enhance customer experience, and to optimize the overall performance. Apache Spark is leveraged at eBay through Hadoop YARN.YARN manages all the cluster resources to run generic tasks. EBay spark users leverage the Hadoop clusters in the range of 2000 nodes, 20,000 cores, and 100TB of RAM through YARN.

Explore Categories

Apache Hive Projects	Apache Hbase Projects
Apache Pig Projects	Hadoop HDFS Projects
Apache Oozie Projects	Apache Impala Projects
Apache Flume Projects	Apache Sqoop Projects
Spark SQL Projects	Spark GraphX Projects
Spark Streaming Projects	Spark MLlib Projects
PySpark Projects	Apache Zepellin Projects	Neo4j Projects

Redis Projects   **Microsoft Azure Projects**

Google Cloud Projects GCP   **AWS Projects**

Spark Use Cases in Healthcare

As healthcare providers look for novel ways to enhance the quality of healthcare, Apache Spark is slowly becoming the heartbeat of many healthcare applications. Many healthcare providers are using Apache Spark to analyse patient records along with past clinical data to identify which patients are likely to face health issues after being discharged from the clinic. This helps hospitals prevent hospital re-admittance as they can deploy home healthcare services to the identified patient, saving on costs for both the hospitals and patients.

Apache Spark is used in genomic sequencing to reduce the time needed to process genome data. Earlier, it took several weeks to organize all the chemical compounds with genes but now with Apache spark on Hadoop it just takes few hours. This use case of spark might not be so real-time like other but renders considerable benefits to researchers over earlier implementation for genomic sequencing.

Companies Using Spark in Healthcare Industry

Apache Spark at MyFitnessPal

The largest health and fitness community MyFitnessPal helps people achieve a healthy lifestyle through better diet and exercise. MyFitnessPal uses apache spark to clean the data entered by users with the end goal of identifying high quality food items. Using Spark, MyFitnessPal has been able to scan through food calorie data of about 80 million users. Earlier, MyFitnessPal used Hadoop to process 2.5TB of data and that took several days to identify any errors or missing information in it.

> Get More Practice, More **Big Data and Analytics Projects**, and More guidance.Fast-Track Your Career Transition with ProjectPro

Spark Use Cases in Media & Entertainment Industry

Apache Spark is used in the gaming industry to identify patterns from the real-time in-game events and respond to them to harvest lucrative business opportunities like targeted advertising, auto adjustment of gaming levels based on complexity, player retention and many more.

Few of the video sharing websites use apache spark along with MongoDB to show relevant advertisements to its users based on the videos they view, share and browse.

Top 15 Latest Recipes

Companies Using Spark in Media & Entertainment Industry

Apache Spark at Yahoo for News Personalization

Yahoo uses Apache Spark for personalizing its news webpages and for targeted advertising. It uses machine learning algorithms that run on Apache Spark to find out what kind of news - users are interested to read and categorizing the news stories to find out what kind of users would be interested in reading each category of news.

Earlier the machine learning algorithm for news personalization required 15000 lines of C++ code but now with Spark Scala the machine learning algorithm for news personalization has just 120 lines of Scala programming code. The algorithm was ready for production use in just 30 minutes of training, on a hundred million datasets.

Apache Spark at Conviva

The largest streaming video company Conviva uses Apache Spark to deliver quality of service to its customers by removing the screen buffering and learning in detail about the network conditions in real-time. This information is stored in the video player to manage live video traffic coming from close to 4 billion video feeds every month, to ensure maximum play-through. Apache Spark is helping Conviva reduce its customer churn to a great extent by providing its customers with a smooth video viewing experience.

Apache Spark at Netflix

Netflix uses Apache Spark for real-time stream processing to provide online recommendations to its customers. Streaming devices at Netflix send events which capture all member activities and play a vital role in personalization. It processes 450 billion events per day which flow to server side applications and are directed to Apache Kafka.

Apache Spark at Pinterest

Pinterest is using apache spark to discover trends in high value user engagement data so that it can react to developing trends in real-time by getting an in-depth understanding of user behaviour on the website.

Get confident to build end-to-end projects.

Access to a curated library of 250+ end-to-end industry projects with solution code, videos and tech support.

Request a demo

Spark Use Cases in Travel Industry

Companies Using Spark in Travel Industry

Apache Spark at TripAdvisor

TripAdvisor, a leading travel website that helps users plan a perfect trip is using Apache Spark to speed up its personalized customer recommendations. TripAdvisor uses apache spark to provide advice to millions of travellers by comparing hundreds of websites to find the best hotel prices for its customers. The time taken to read and process the reviews of the hotels in a readable format is done with the help of Apache Spark.

Apache Spark at OpenTable

OpenTable, an online real time reservation service, with about 31000 restaurants and 15 million diners a month, uses Spark for training its recommendation algorithms and for NLP of the restaurant reviews to generate new topic models. OpenTable has achieved 10 times speed enhancements by using Apache Spark. Spark has helped reduce the run time of machine learning algorithms from few weeks to just a few hours resulting in improved team productivity.

The spike in increasing number of spark use cases is just in its commencement and 2016 will make Apache Spark the big data darling of many other companies, as they start using Spark to make prompt decisions based on real-time processing through spark streaming. These are just some of the use cases of the Apache Spark ecosystem. If you know any other companies using Spark for real-time processing, feel free to share with the community, in the comments below.

Build an Awesome Job Winning Project Portfolio with Solved End-to-End Big Data Projects

Spark Use Cases in Gaming Industry

Apache Spark is used in the gaming industry to identify patterns from real-time in-game events. It helps companies to harvest lucrative business opportunities like targeted advertising, auto adjustment of gaming levels based on complexity. It also provides in-game monitoring, player retention, detailed insights, and many more.

Companies Using Spark in Gaming Industry

Riot Games

Game developers have to manage everything from performance to in-game abuses. Spark improves the gaming experience of the users, it also helps in processing different game skins, different game characters, in-game points,

and much more. It helps with performance improvement, offers, and efficiency. Riot can now detect the cause which made the game slow and laggy, so they can solve problems on time without impacting users.

Riot Games uses Apache Spark to minimize the in-game toxicity. Whether you are winning or losing, some players get into a rage. Game developers at Riot use Spark MLlib to train their models on NLP for words, short forms, initials, etc. to understand how a player interacts and they can even disable their account if required.

Tencent

Tencent has the biggest market of mobile gaming users base, similar to riot it develops multiplayer games. Tencent uses spark for its in-memory computing feature that boosts data processing performance in real-time in a big data context while also assuring fault tolerance and scalability. It uses Apache Spark to analyze multiplayer chat data to reduce the usage of abusive languages in-game chat.

Spark Use Cases in Software & Information Service Industry

Spark use cases in Computer Software and Information Technology and Services takes about **32%** and **14%** respectively in the global market. Apache Spark is designed for interactive queries on large datasets; its main use is streaming data which can be read from sources like Kafka or Hadoop output or even files on disk. Apache Spark also has a wide range of built-in computational engines such as SQL and Streaming algorithms that can be used to perform computations on its data sets There are many interesting properties that make Apache Spark attractive to use for streaming data analysis.

Spark in Software & Information Service Industry

Databricks

Databricks was developed by creators of spark. It is a [cloud-optimized platform](#) to run Spark and ML applications on AWS and Azure, also a comprehensive training program. They are working on spark to expand the project and make new progress to it. The company has also developed various open-source applications like Delta Lake, MLflow, and Koalas, popular open-source projects that span data engineering, data science, and machine learning.

Hearst

It is a leading global media information and services company. Its main goal is to provide services to many major businesses, from television channels to financial services. Using Apache Spark Streaming Hearst's team gleans real-time insights on articles/news items performing well and identifies content that is trending.

FINRA

FINRA is a Financial Services company that helps get real-time data insights of billions of data events. Using Apache Spark, it can test things on real data from

the market, improving its ability to provide investor security and promote

market integrity.

Big Data Analytics Projects using Spark- Spark Projects

Spark project 1: Create a data pipeline based on messaging using Spark and Hive

Problem: A data pipeline is used to transport data from source to destination through a series of processing steps. The data source could be other databases, api's, json format, csv files etc. Final destination could be another process or visualization tools. In between this, data is transformed into a more intelligent and readable format.

Technologies used: AWS, Spark, Hive, Scala, Airflow, Kafka.

Solution Architecture: This implementation has the following steps: Writing events in the context of a data pipeline. Then designing a data pipeline based on messaging. This is followed by executing the file pipeline utility. After this we load data from a remote URL, perform Spark transformations on this data before moving it to a table. Then Hive is used for data access.

Spark Project 2: Building a Data Warehouse using Spark on Hive

Problem: Large companies usually have multiple storehouses of data. All this data must be moved to a single location to make it easy to generate reports. A data warehouse is that single location.

Technologies used:HDFS, Hive, Sqoop, Databricks Spark, Dataframes.

Solution Architecture: In the first layer of this spark project first moves data to hdfs. The hive tables are built on top of hdfs. Data comes through batch processing. Sqoop is used to ingest this data. Dataframes are used to store instead of RDD. In the 2nd layer, we normalize and denormalize the data tables. Then transformation is done using Spark Sql. This transformed data is moved to HDFS. In the final 3rd layer visualization is done.

Most Watched Projects

Hands-On Real Time PySpark Project for Beginners	NLP Project to Build a Resume Parser in Python using Spacy	Linear Regr Model Proj Python for Part 1
View Project	View Project	View P

View all Most Watched Projects

Spark Use Cases in Advertising

With the increased usage of digital and social media adoption, Apache Spark is helping companies achieve their business goals in various ways. It helps to compute additional data that enrich a dataset. Broadly, this includes gathering metadata about the original data and computing probability distributions for categorical features. It can be used to add additional fields like the mode or median of numerical values in categorical columns like "color" or "age". It can also be used to fill in missing features based on other similar tuples in the same table such as zip_code, gender, or state_province. It is used by advertisers to combine all sorts of data and provide user-based and targeted ads.

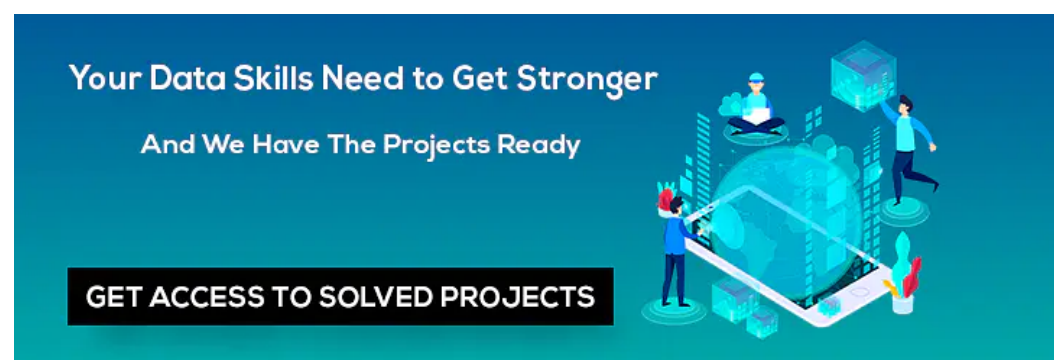## Companies Using Spark in Advertising Industry

Yelp

Founded in 2004, Yelp helps connect people with local businesses. From booking a table to getting food delivered. The advertising targeting team at Yelp uses prediction algorithms to figure out how likely it is for a person to interact with an ad. Yelp enhanced revenue and ad click-through rate by utilizing Apache Spark on Amazon EMR to analyze enormous volumes of data and train machine learning models.

Gumgum

It is an AI-focused technology and digital media company. They have been using machine learning to extract value from digital content for over a long time. It is an in-image and in-screen advertising platform, employing Spark on Amazon EMR. For forecasting, log processing, ad hoc analysis, and a lot more. Spark's speed helps gumgum save lots of time and resources. It uses computer vision and NLP to identify and score different types of content

Recommended Reading:

- **How to Become a Big Data Engineer in 2021**
- **Big Data Engineer Salary - How Much Can You Make in 2021?**

PREVIOUS　　　　　　　　　　　　　　　　　　　　　　　NEXT

Trending Project Categories

Machine Learning Projects

Trending Projects

Walmart Sales Forecasting Data Science Project

BigMart Sales Prediction ML Project

Trending Blogs

Machine Learning Projects for Beginners with Source Code

Data Science Projects for Beginners with

Data Science Projects

Deep Learning Projects

Big Data Projects

Apache Hadoop Projects

Apache Spark Projects

Show more

Music Recommender System Project

Credit Card Fraud Detection Using Machine Learning

Resume Parser Python Project for Data Science

Time Series Forecasting Projects

Show more

Source Code

Big Data Projects for Beginners with Source Code

IoT Projects for Beginners with Source Code

Data Analyst vs Data Scientist

Data Science Interview Questions and Answers

Show more

Trending Recipes

Search for a Value in Pandas DataFrame

Pandas Create New Column based on Multiple Condition

LSTM vs GRU

Plot ROC Curve in Python

Python Upload File to Google Drive

Optimize Logistic Regression Hyper Parameters

Show more

Trending Tutorials

PCA in Machine Learning Tutorial

PySpark Tutorial

Hive Commands Tutorial

MapReduce in Hadoop Tutorial

Apache Hive Tutorial -Tables

Linear Regression Tutorial

Show more

ProjectPro

© 2022 Iconiq Inc.

About us

Contact us

Privacy policy

User policy

Write for ProjectPro

Data Science Projects

Deep Learning Projects

Big Data Projects

Apache Hadoop Projects

Apache Spark Projects

Show more

Music Recommender System Project

Credit Card Fraud Detection Using Machine Learning

Resume Parser Python Project for Data Science

Time Series Forecasting Projects

Show more

Source Code

Big Data Projects for Beginners with Source Code

IoT Projects for Beginners with Source Code

Data Analyst vs Data Scientist

Data Science Interview Questions and Answers