Table Of Contents:

1. Advancement of Data Science
2. Comparison of Spark 2.4 and 3.0
3. Machine Learning Implementation (250 words)
4. Data set
5. Collaborative filtering:  
    Features of the model, key parameters and configuration Evaluation
6. Logistic regression:  
    Features of the model, key parameters and configuration Evaluation
7. Advancement in Data Science:

The idea of combining applied statistics and computer science gave birth to data science. Scientists found they could utilize data to address real-world problems and produce accurate fact-based forecasts in addition to collecting data and solving statistical difficulties. Deep learning, natural language processing, and computer vision are examples of technologies that have evolved as a result of Data Science's growth in the field of research and practical application throughout the last century(‘Top 10 AI and Data Science Trends in 2022’, 2022, p. 10). In the last 2 years, Machine learning has risen in popularity as a result of advancements in programming languages like Python, data gathering, technology, and large data generation around the world and we're seeing greater advances in AI and machine learning. As a result of the influx of new data, organizations seeking new ways to maximize profit and make better decisions. Data science has also begun to spread in other areas, including medical, engineering, and more. In the not-too-distant future, we may witness a period of widespread labor automation. As a result, data scientists are working nonstop to improve deep learning to make machines smarter. It has supported the development of machine learning (ML) as a means of reaching artificial intelligence (AI), a field of technology that is rapidly transforming the way we work and live in general. These advancements could lead to advanced robotics combined with a formidable AI.(“Top Data Science & AI Trends For 2022,” n.d.)  
The following are some of the most significant data science breakthroughs in the last two years:

* **AI and Data Solutions on Cloud**: Large amounts of data are already being produced. The problem is gathering, labeling, cleaning, organizing, structuring, and analyzing such a large amount of data in one place. The answer is a cloud-based platform.

## Low code technology and high models have been improved: Companies are beginning to leverage out-of-the-box foundation models as they begin to apply AI in the industry, shortening time-to-value for AI solutions in areas like language, vision, and more. Artificial intelligence (AI) will have a tremendous impact on citizen development.

## Analytics with Augmented Data: Augmented analytics is a sort of data analytics that uses AI, machine learning, with natural language processing to automate the analysis of massive amounts of data. By assisting with data preparation, data processing, analytics, and visualization, AI, ML, and NLP enable specialists to study data and deliver in-depth analyses and projections. Data from both inside and outside the firm can be integrated using augmented analytics. AI and machine learning capabilities have been progressively and directly deployed inside analytics and BI systems to aid business users rather than just data specialists, thanks to the emergence of visual-based data discovery tools in recent years.(“The Past, Present, and Future of Data Science,” 2021)

## AutoML: Automated machine learning is a strategy for applying machine learning (ML) models to real-world scenarios using automation (AutoML).

## Edge computing: Edge computing or edge intelligence, is the processing and aggregation of data at the network's edge.

## Filtering Data Automatically: If the data isn’t filtered enough for analytics, it's useless. It also refers to erroneous data, data redundancy, and data duplication with no structure or format. Automated data cleansing and scrubbing solutions to improve data analytics gain more trustworthy insights for analyzing huge data.

## Data Science and Blockchain: Data analysis is still a time-consuming process that requires the efforts of data scientists and blockchain technology can effectively tackle the problem.(“Is Data Science Still a Rising Career in 2021 | by Chris Zaire | Towards Data Science,” n.d.)

## Comparison of Spark 2.4 and 3.0

## Apache Spark 2.4.0: It’s the fifth release in the 2.x line. This release adds Barrier Execution Mode for improved deep learning framework interaction, introduces 30+ built-in and higher-order functions to make dealing with complicated data types easier, and improves the K8s integration. Major features:

## Barrier Execution Mode: To better interface with deep learning frameworks, support Barrier Execution Mode in the scheduler.

## Scala 2.12 Support: Scala 2.12 support is now experimental. Scala 2.12 may now be used to construct Spark and to write Spark apps.

## Higher-order functions: Many new built-in functions, including higher-order functions, have been added to make working with complicated data types easier.

Apache Spark 3.0.0: It is the first release of the 3.x line. This release is based on git tag v3.0.0. It expands on many of the innovations from Spark 2.x, introducing new ideas while also continuing long-term projects in development. In TPC-DS 30TB benchmark, Spark 3.0 is nearly 2 times faster than Spark 2.4. Major features:

* **Adaptive Query Execution**: One of the most important aspects of Spark 3.0 is Adaptive Query Execution (AQE), which reoptimizes and modifies query plans based on runtime statistics acquired during query execution.
* **Upgrades to the language version**: Spark 3.0 was published, and it now supports more languages. Spark now supports Python 3 (Python 2.x), Scala 2.12, and JDK 11 as language versions.
* **Structure streaming has a new UI**: Spark Web UI now has a new Structured Streaming tab to monitor Structured streaming applications in version 3.0.

1. Machine Learning Implementation
2. **Data Set**: This dataset describes a 5-star rating from Movie Lens, a movie recommendation service. It contains 100836 ratings and 3683 tag applications across 9742 movies. These data were created by 610 users.(“Movie Lens Dataset,” n.d.)
3. **Collaborative Filtering**: Collaborative filtering is a technique for creating automatic predictions (filtering) about a user's interests by collecting preferences or taste information from a large number of people. Alternating Least Square (ALS) is a matrix factorization algorithm, implemented in Apache Spark ML and built for large-scale Collaborative filtering problems(“Collaborative Filtering - Spark 2.2.0 Documentation,” n.d.). Its key parameters are:
   1. maxIter: the maximum number of iterations to run (set to 20)
   2. rank: the number of latent factors in the model (defaults to 10)
   3. regParam: the regularization parameter in ALS (set to 0.05)

The model was evaluated on 20% of total data and mse was found out to be 0.891.

1. **Logistic Regression**: PySpark Logistic Regression is a classification that predicts the dependency of data over each other in the PySpark ML model. It is a faster way of classification of data and works fine with larger data set with accurate results. Its key parameters are:
   1. maxIter: the maximum number of iterations to run (set to 20)
   2. regParam: the regularization parameter (set to 0.05)
   3. elasticNetParam: the combination of L1 and L2 regularizer (set to 0.8)

The modal was evaluated for both 80% training dataset and 20%test dataset and the rmse was found to be 3.6521 and 3.6585 respectively(“Classification and regression - Spark 3.2.1 Documentation,” n.d.)

## References

Classification and regression - Spark 3.2.1 Documentation [WWW Document], n.d. URL https://spark.apache.org/docs/latest/ml-classification-regression.html (accessed 2.9.22).

Collaborative Filtering - Spark 2.2.0 Documentation [WWW Document], n.d. URL https://spark.apache.org/docs/2.2.0/ml-collaborative-filtering.html (accessed 2.9.22).

Is Data Science Still a Rising Career in 2021 | by Chris Zaire | Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/is-data-science-still-a-rising-career-in-2021-722281f7074c (accessed 2.9.22).

Movie Lens Dataset [WWW Document], n.d. URL https://kaggle.com/aigamer/movie-lens-dataset (accessed 2.9.22).

The Past, Present, and Future of Data Science [WWW Document], 2021. . Dataquest. URL https://www.dataquest.io/blog/evolution-of-data-science-growth-innovation/ (accessed 2.9.22).

Top 10 AI and Data Science Trends in 2022, 2022. . Analytics Vidhya. URL https://www.analyticsvidhya.com/blog/2022/02/top-10-ai-and-data-science-trends-in-2022/ (accessed 2.9.22).

Top Data Science & AI Trends For 2022 [WWW Document], n.d. URL https://analyticsindiamag.com/top-data-science-ai-trends-for-2022/ (accessed 2.9.22).