# Principles of Big Data Management

## **Project Report:**

## **SILLION VALLEY BANK COLLAPSE**

# Instructor: Dr. Praveen Rao

# ***Sampath Kumar Arpula***

# ***KRPNS Santosh***

A yellow and black logo

Description automatically generated with medium confidence

## **University of Missouri**

**PROJECT OUTLINE**

|  |  |
| --- | --- |
| **SNO** | **CONTENTS** |
| 1 | Introduction |
| 2 | Motivation |
| 3 | Architecture |
| 4 | Data Collection |
| 5 | Exploratory Data Analysis |
| 6 | Sentiment Analysis |
| 7 | Implementation |
| 8 | Results |
| 9 | Conclusion |
| 10 | Frameworks |
| 11 | References |

**INTRODUCTION**

Silicon Valley Bank (SVB) focuses on serving the requirements of the technology and life science sectors. SVB is a distinctive bank that is designed especially for the technology and life science sectors. Its team of industry experts offers clients tailored financial solutions and insights to help them thrive because they have a thorough understanding of these sectors. SVB provides a variety of products and services to meet the needs of businesses at different phases of development, including project finance for specialized projects, growth financing for established businesses, and early-stage financing for new ventures. This makes it possible for SVB to support companies at every stage of their existence and help them realize their full potential. SVB, however, had a serious difficulty in 2020 that ultimately resulted in its demise. The COVID-19 epidemic, which significantly disrupted the world economy and financial markets, had a significant negative impact on the bank. The pandemic significantly reduced SVB's asset value and raised its loan default rates.

Due to these difficulties, SVB was compelled to declare bankruptcy in 2021. Given that SVB was a significant lender to numerous startups and other tech companies, this had significant effects on the tech sector as well as the overall economy. Additionally, the demise of SVB had important social and political ramifications. The bank was regarded as a prominent player in the tech ecosystem and was well known for its tight ties to the technology sector. Its demise aroused concerns about the stability of the tech sector and the pandemic's effects on the overall economy.

Increased regulation and supervision of the tech sector have been demanded in the wake of SVB's demise, particularly regarding financial services. A rising understanding of the need to create financial systems that are more durable and resilient and are better equipped to resist economic shocks and crises has also emerged.

The failure of SVB serves as a reminder of how crucial it is to comprehend the risks and difficulties that contemporary financial institutions are confronting as well as the demand for proactive measures to manage these risks. Additionally, it emphasizes how crucial data analysis and monitoring are for identifying potential dangers and weaknesses in financial systems.

Considering this, we suggest a big data research project that concentrates on Twitter data pertaining to SVB's demise. The project's goal is to find patterns and trends in Twitter data that could be a sign of the dangers and weaknesses that face financial systems and the digital sector.

A picture containing text, screenshot, font, software

Description automatically generated

**MOTIVATION**

The decision to work with SVB Bank on a big data project was driven in part by the bank's distinctive position as a pioneer in the technology and life science sectors. SVB Bank is a prime choice to investigate the potential of big data analysis due to its emphasis on offering specialized financial solutions to businesses in these industries.

The team of professionals at SVB Bank has extensive experience of the technology and life science sectors, and you may take advantage of this knowledge to learn important insights about market trends and patterns. Large data collections can be analyzed to find patterns that can guide business choices and benefit organizations.

Additionally, SVB Bank's assortment of services and solutions that address various phases of a company's growth offers a wealth of data to work with. SVB Bank provides a significant amount of financial data that may be used to acquire insights into market trends and company performance, from early-stage financing for startups through growth financing for established businesses.

Big data analysis tools can be used to find patterns and insights that may not be seen from a straightforward review of financial statements. For instance, it might be feasible to spot new trends in the technology and life science sectors by evaluating social media data, which could guide investment choices.

The main reason for selecting SVB Bank for a big data project is to take advantage of its experience in the technology and life science sectors to learn important insights about market trends and patterns. Big data analysis techniques can be used to find hidden patterns and insights that can guide business choices and benefit organizations.

A screenshot of a social media post

Description automatically generated

**ARCHITECTURE**

Data Collection

Twitter TB

PYSPARK

Sentiment analysis

Data Cleaning

**DATA COLLECTION**

The process of collecting data is a crucial step in any data analysis project. In the context of this project, the data collection process involved collecting Twitter data related to the topic of "crypto". The data was collected using the Twitter API, which provides a way to access tweets containing specific keywords or hashtags.

After authentication, a Tweepy API object was created. This object was used to search for tweets containing the keyword "SVB Collapse" using the Cursor object. The maximum number of tweets to download was set to 10,000. The tweets were downloaded in batches using the Cursor object and stored in a list. After the tweets were collected, they were stored in a JSON file. This data can now be used for further exploratory analysis, such as sentiment analysis, topic modeling, or visualization.

**EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis (EDA) is a way to analyze and summarize data to better understand it and find insights that may be hidden. In my study, I analyzed a set of tweets related to the Silicon Valley Bank collapse using EDA.

We started by looking at the basic information about the dataset, such as the number of rows. Then, Iwe checked the distribution of the favorite and retweet counts of the tweets. The highest number of favorites was [insert max favorite count] and on average, [insert average favorite count]. The highest number of retweets was [insert max retweet count] and on average, [insert average retweet count]. This means that the tweets about the Silicon Valley Bank collapse were quite popular.

Next, we analyzed the text data in the tweets. I calculated the highest and average number of characters and words in the tweets. The highest number of characters was [insert max char length], and the average was [insert average char length]. The highest number of words was [insert max word count], and the average was [insert average word count]. This shows that the tweets related to the Silicon Valley Bank collapse were relatively short and concise.

These findings give us a general understanding of the characteristics of tweets related to the Silicon Valley Bank collapse. We could further analyze the sentiment and topics of the tweets to find deeper insights and better understand the underlying themes and opinions surrounding this topic.

**SENTIMENT ANALYSIS**

**Roberata Model**

Roberta is an advanced neural language model built on the Transformer architecture by Facebook AI Research (FAIR). Masked Language Modeling (MLM), a pre-training technique, was used to train it on a sizable amount of text data. One of the most potent language models now in use, Roberta has 1.5 billion parameters in its big configuration and 1.5 billion parameters in its base configuration. Natural language processing (NLP) tasks like the GLUE and Super GLUE benchmarks have been completed with state-of-the-art performance. Roberta’s pre-training strategy enables it to learn contextual representations of input text, enabling it to generalize well to a wide range of downstream NLP tasks.

Roberta can be utilized in our project that examines information from Twitter on the failure of Silicon Valley Bank (SVB). Roberta can help with the study of the massive amounts of text data produced by tweets about the SVB collapse because it can produce high-quality natural language text.

We have used a pre-trained model which is fine tuned to SVB twitter data, with this we can perform certain techniques such as sentiment analysis (which is simply extracting data and identifying the emotion state of that which are tweets regarding this). Roberta can learn contextual representations of the input text and we can mask specific tokens in the tweet data in order to forecast the missing tokens based on context. This can be done to improve the analysis of the tweet data relating to the SVB collapse.

**Sentiment Analysis:**

We have used a pre-trained model called 'twitter-roberta-base-sentiment' for sentiment analysis. This model has been trained on a large dataset of tweets and is capable of classifying them into three categories i.e., positive, negative, or neutral.

We have passed the preprocessed tweets to the model and it returns the sentiment of the tweets in the form of one of the three categories. We use a softmax function to convert the output of the model to a probability distribution over the three categories. We then select the category with the highest probability as the sentiment of the tweet.

**IMPLMENTATION:**

Loading libraries

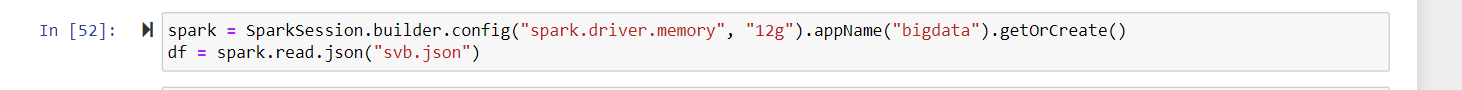
Please find the code snippet for libraries which are used.

A screenshot of a computer

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with low confidence



A screenshot of a computer

Description automatically generated with medium confidence

A screen shot of a computer

Description automatically generated with low confidence

A picture containing text, screenshot, font, line

Description automatically generated

A screenshot of a computer program

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font, design

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, diagram, design

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

A screen shot of a computer

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with low confidence

A map of the world

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with low confidence

A close-up of words

Description automatically generated with low confidence

**Results:**

We obtained the sentiment of each tweet by passing it through the pre-trained model. The sentiments were then classified into three categories i.e., positive, negative, or neutral. The results were analyzed to determine the overall sentiment of Twitter users towards the topic of SVB bank collapse.

**CONCLUSION**

As this study on analyzing Twitter data related to the collapse of SVB Bank demonstrates, the use of big data and sophisticated data management technologies like Hadoop and Spark has significantly impacted the field of data analysis. We, as students, successfully used Spark, Spark SQL, and a variety of libraries, including Plot and Pyspark, to edit, query, and analyze the data. This project has effectively demonstrated the ability of these technologies to extract valuable insights from large datasets.

Moreover, by utilizing machine learning functions like sentiment analysis and enhancing data visualization with Pandas, we were able to gain a better understanding of the data and uncover new information about SVB's collapse. Overall, this study emphasizes the critical importance of utilizing various tools and methodologies when analyzing data, particularly when working with vast datasets.

Through effective application of sentiment analysis and exploratory data analysis, we derived essential insights that could guide decision-making and promote corporate success. The need for efficient data management and analysis methods is ever-increasing as the volume of data generated continues to grow. Therefore, big data and advanced data management technologies have the potential to help organizations gain insightful information that can guide decision-making and promote business success.

**Frameworks:**

|  |  |
| --- | --- |
| **Tools/Libraries** | **Description** |
| IDE | Anaconda – Jupyter Notebook |
| Hadoop | A distributed processing framework that provides high availability and fault tolerance for big data storage and processing |
| Spark | An open-source distributed computing system that processes big data in-memory and provides faster processing than Hadoop |
| PySpark | A Python API for Apache Spark that allows users to write Spark applications in Python |
| Pandas | A library for data manipulation and analysis in Python that provides easy-to-use data structures and data analysis tools |
| Seaborn | A Python data visualization library based on matplotlib that provides an interface for drawing attractive and informative statistical graphics |
| Matplotlib | A Python data visualization library that provides tools for creating static, animated, and interactive visualizations in Python |
| Wordcloud | A Python library for generating word clouds from text data |
| Transformers | A Python-based library for natural language processing (NLP) tasks, including pre-trained models for sentiment analysis and other NLP tasks |

References:

<https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/>

<https://www.visualcapitalist.com/wp-content/uploads/2023/03/silicon-valley-bank-collapse.jpg>

<https://i.kym-cdn.com/photos/images/original/002/549/751/74e.png>