BIG DATA ANALYTICS

UNIT - II

Stream Processing: Mining data streams: Introduction to Streams Concepts, Stream Data Model and Architecture, Stream Computing, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Counting Oneness in a Window, Decaying Window, Real time Analytics Platform (RTAP) Applications, Case Studies - Real Time Sentiment Analysis - Stock Market Predictions.

Stream Processing

Stream processing is a method of data processing that involves continuously processing data in real-time as it is generated, rather than processing it in batches. In stream processing, data is processed incrementally and in small chunks as it arrives, making it possible to analyze and act on data in real-time.

Stream processing is particularly useful in scenarios where data is generated rapidly, such as in the case of IoT devices or financial markets, where it is important to detect anomalies or patterns in data quickly. Stream processing can also be used for real-time data analytics, machine learning, and other applications where real-time data processing is required.

There are several popular stream processing frameworks, including Apache Flink, Apache Kafka, Apache Storm, and Apache Spark Streaming. These frameworks provide tools for building and deploying stream processing pipelines, and they can handle large volumes of data with low latency and high throughput.

Mining data streams

Mining data streams refers to the process of extracting useful insights and patterns from continuous and rapidly changing data streams in real-time. Data streams are typically high-volume and high-velocity, making it challenging to analyze them using traditional data mining techniques.

Mining data streams requires specialized algorithms that can handle the dynamic nature of data streams, as well as the need for real-time processing. These algorithms typically use techniques such as sliding windows, online learning, and incremental processing to adapt to changing data patterns over time.

Applications of mining data streams include fraud detection, network intrusion detection, predictive maintenance, and real-time recommendation systems. Some popular algorithms for mining data streams include Frequent Pattern Mining (FPM), clustering, decision trees, and neural networks.

Mining data streams also requires careful consideration of the computational resources required to process the data in real-time. As a result, many mining data stream algorithms are designed to work with limited memory and processing power, making them well-suited for deployment on edge devices or in cloud-based architectures.

Introduction to Streams Concepts

In computer science, a stream refers to a sequence of data elements that are continuously generated or received over time. Streams can be used to represent a wide range of data, including audio and video feeds, sensor data, and network packets.

Streams can be thought of as a flow of data that can be processed in real-time, rather than being stored and processed at a later time. This allows for more efficient processing of large volumes of data and enables applications that require real-time processing and analysis.

Some important concepts related to streams include:

1. **Data Source:** A stream's data source is the place where the data is generated or received. This can include sensors, databases, network connections, or other sources.
2. **Data Sink:** A stream's data sink is the place where the data is consumed or stored. This can include databases, data lakes, visualization tools, or other destinations.
3. **Streaming Data Processing:** This refers to the process of continuously processing data as it arrives in a stream. This can involve filtering, aggregation, transformation, or analysis of the data.
4. **Stream Processing Frameworks:** These are software tools that provide an environment for building and deploying stream processing applications. Popular stream processing frameworks include Apache Flink, Apache Kafka, and Apache Spark Streaming.
5. **Real-time Data Processing:** This refers to the ability to process data as soon as it is generated or received. Real-time data processing is often used in applications that require immediate action, such as fraud detection or monitoring of critical systems.

Overall, streams are a powerful tool for processing and analyzing large volumes of data in real-time, enabling a wide range of applications in fields such as finance, healthcare, and the Internet of Things.

Stream Data Model and Architecture

Stream data model is a data model used to represent the continuous flow of data in a stream processing system. The stream data model typically consists of a series of events, which are individual pieces of data that are generated by a data source and processed by a stream processing system.

The architecture of a stream processing system typically involves three main components: data sources, stream processing engines, and data sinks.

1. **Data sources:** The data sources are the components that generate the events that make up the stream. These can include sensors, log files, databases, and other data sources.
2. **Stream processing engines:** The stream processing engines are the components responsible for processing the data in real-time. These engines typically use a variety of algorithms and techniques to filter, transform, aggregate, and analyze the stream of events.
3. **Data sinks:** The data sinks are the components that receive the output of the stream processing engines. These can include databases, data lakes, visualization tools, and other data destinations.

The architecture of a stream processing system can be distributed or centralized, depending on the requirements of the application. In a distributed architecture, the stream processing engines are distributed across multiple nodes, allowing for increased scalability and fault tolerance. In a centralized architecture, the stream processing engines are run on a single node, which can simplify deployment and management.

Some popular stream processing frameworks and architectures include Apache Flink, Apache Kafka, and Lambda Architecture. These frameworks provide tools and components for building scalable and fault-tolerant stream processing systems, and can be used in a wide range of applications, from real-time analytics to internet of things (IoT) data processing.

Stream Computing

Stream computing is the process of computing and analyzing data streams in real-time. It involves continuously processing data as it is generated, rather than processing it in batches. Stream computing is particularly useful for scenarios where data is generated rapidly and needs to be analyzed quickly.

Stream computing involves a set of techniques and tools for processing and analyzing data streams, including:

1. **Stream processing frameworks:** These are software tools that provide an environment for building and deploying stream processing applications. Popular stream processing frameworks include Apache Flink, Apache Kafka, and Apache Storm.
2. **Stream processing algorithms:** These are specialized algorithms that are designed to handle the dynamic and rapidly changing nature of data streams. These algorithms use techniques such as sliding windows, online learning, and incremental processing to adapt to changing data patterns over time.
3. **Real-time data analytics:** This involves using stream computing techniques to perform real-time analysis of data streams, such as detecting anomalies, predicting future trends, and identifying patterns.
4. **Machine learning:** Machine learning algorithms can also be used in stream computing to continuously learn from the data stream and make predictions in real-time.

Stream computing is becoming increasingly important in fields such as finance, healthcare, and the Internet of Things (IoT), where large volumes of data are generated and need to be processed and analyzed in real-time. It enables businesses and organizations to make more informed decisions based on real-time insights, leading to better operational efficiency and improved customer experiences.

Sampling Data in a Stream

Sampling data in a stream refers to the process of selecting a subset of data points from a continuous and rapidly changing data stream for analysis. Sampling is a useful technique for processing data streams when it is not feasible or necessary to process all data points in real-time.

There are various sampling techniques that can be used for stream data, including:

1. **Random sampling:** This involves selecting data points from the stream at random intervals. Random sampling can be used to obtain a representative sample of the entire stream.
2. **Systematic sampling:** This involves selecting data points at regular intervals, such as every tenth or hundredth data point. Systematic sampling can be useful when the stream has a regular pattern or periodicity.
3. **Cluster sampling:** This involves dividing the stream into clusters and selecting data points from each cluster. Cluster sampling can be useful when there are multiple sub-groups within the stream.
4. **Stratified sampling:** This involves dividing the stream into strata or sub-groups based on some characteristic, such as location or time of day. Stratified sampling can be useful when there are significant differences between the sub-groups.

When sampling data in a stream, it is important to ensure that the sample is representative of the entire stream. This can be achieved by selecting a sample size that is large enough to capture the variability of the stream and by using appropriate sampling techniques.

Sampling data in a stream can be used in various applications, such as monitoring and quality control, statistical analysis, and machine learning. By reducing the amount of data that needs to be processed in real-time, sampling can help improve the efficiency and scalability of stream processing systems.

Filtering Streams

Filtering streams refers to the process of selecting a subset of data from a data stream based on certain criteria. This process is often used in stream processing systems to reduce the amount of data that needs to be processed and to focus on the relevant data.

There are various filtering techniques that can be used for stream data, including:

1. **Simple filtering:** This involves selecting data points from the stream that meet a specific condition, such as a range of values, a specific text string, or a certain timestamp.
2. **Complex filtering:** This involves selecting data points from the stream based on multiple criteria or complex logic. Complex filtering can involve combining multiple conditions using Boolean operators such as AND, OR, and NOT.
3. **Machine learning-based filtering:** This involves using machine learning algorithms to automatically classify data points in the stream based on past observations. This can be useful in applications such as anomaly detection or predictive maintenance.

When filtering streams, it is important to consider the trade-off between the amount of data being filtered and the accuracy of the filtering process. Too much filtering can result in valuable data being discarded, while too little filtering can result in a large volume of irrelevant data being processed.

Filtering streams can be useful in various applications, such as monitoring and surveillance, real-time analytics, and Internet of Things (IoT) data processing. By reducing the amount of data that needs to be processed and analyzed in real-time, filtering can help improve the efficiency and scalability of stream processing systems.

Counting Distinct Elements in a Stream

Counting distinct elements in a stream refers to the process of counting the number of unique items in a continuous and rapidly changing data stream. This is an important operation in stream processing because it can help detect anomalies, identify trends, and provide insights into the data stream.

There are various techniques for counting distinct elements in a stream, including:

1. **Exact counting:** This involves storing all the distinct elements seen so far in a data structure such as a hash table or a bloom filter. When a new element is encountered, it is checked against the data structure to determine if it is a new distinct element.
2. **Approximate counting:** This involves using probabilistic algorithms such as the Flajolet-Martin algorithm or the HyperLogLog algorithm to estimate the number of distinct elements in a data stream. These algorithms use a small amount of memory to provide an approximate count with a known level of accuracy.
3. **Sampling:** This involves selecting a subset of the data stream and counting the distinct elements in the sample. This can be useful when the data stream is too large to be processed in real-time or when exact or approximate counting techniques are not feasible.

Counting distinct elements in a stream can be useful in various applications, such as social media analytics, fraud detection, and network traffic monitoring. By providing real-time insights into the data stream, counting distinct elements can help businesses and organizations make more informed decisions and improve operational efficiency.

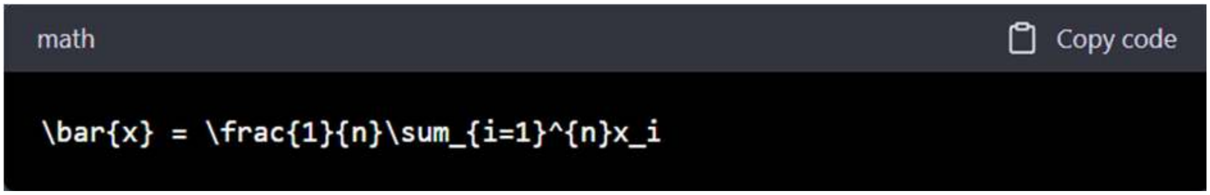
Estimating Moments

In statistics, moments are numerical measures that describe the shape, central tendency, and variability of a probability distribution. They are calculated as functions of the random variables of the distribution, and they can provide useful insights into the underlying properties of the data.

There are different types of moments, but two of the most commonly used are the mean (the first moment) and the variance (the second moment). The mean represents the central tendency of the data, while the variance measures its spread or variability.

To estimate the moments of a distribution from a sample of data, you can use the following formulas:

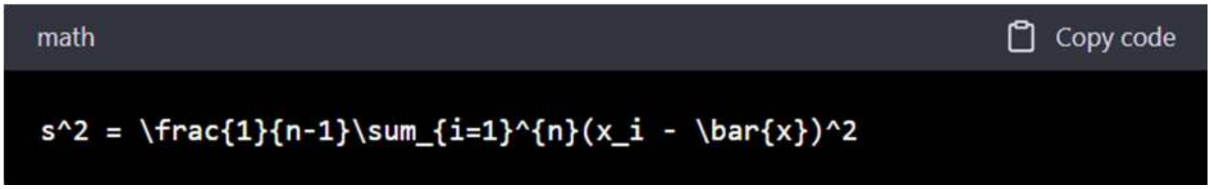
Sample mean (first moment):



```
\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i
```

where n is the sample size, and x_i are the individual observations.

Sample variance (second moment):



```
s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2
```

where n is the sample size, x_i are the individual observations, and s^2 is the sample variance.

These formulas provide estimates of the population moments based on the sample data. The larger the sample size, the more accurate the estimates will be. However, it's important to note that these formulas only work for certain types of distributions (e.g., normal distribution), and for other types of distributions, different formulas may be required.

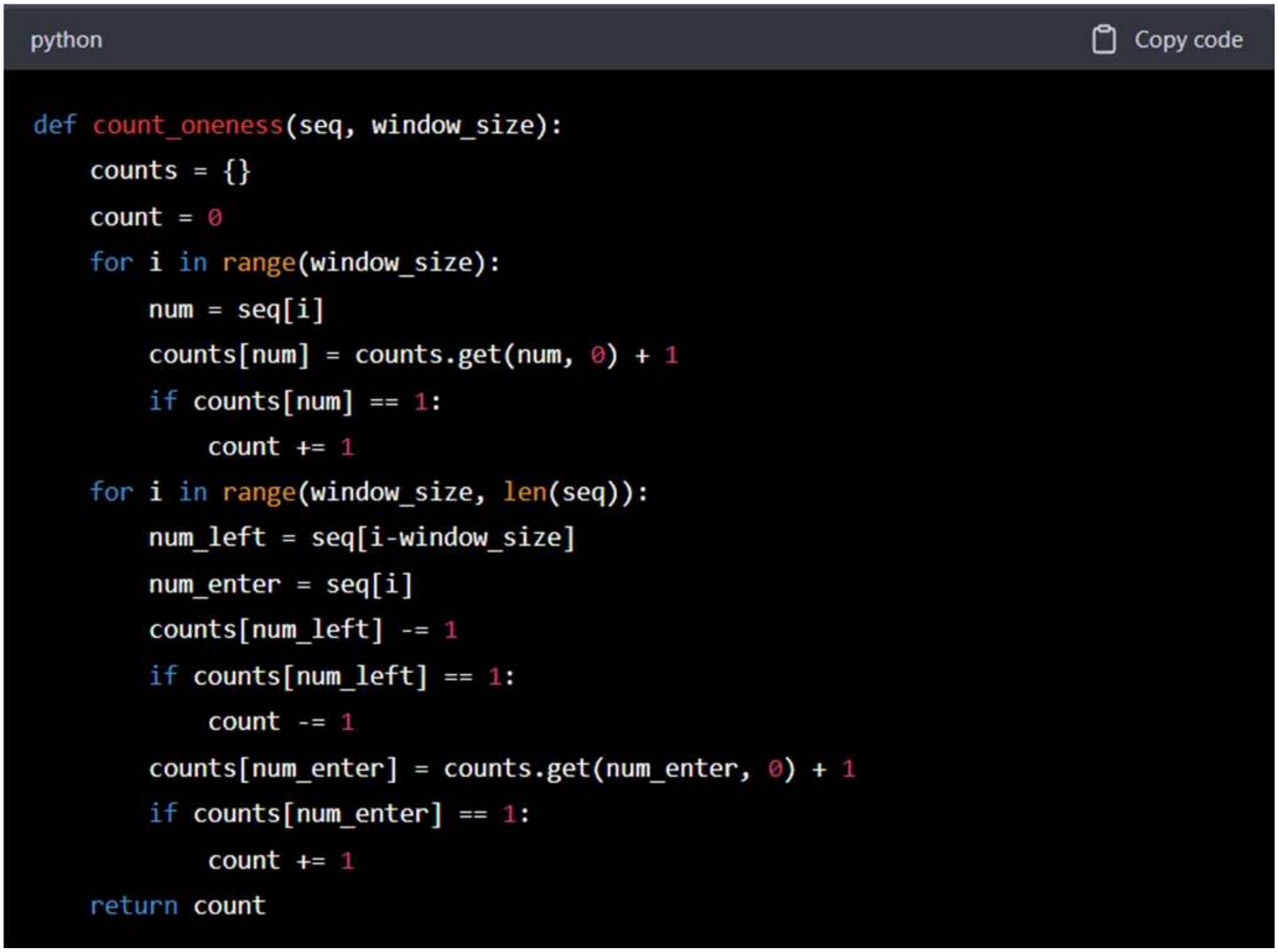
Counting Oneness in a Window

Counting the number of times a number appears exactly once (oneness) in a window of a given size in a sequence is a common problem in computer science and data analysis. Here's one way you could approach this problem:

1. Initialize a dictionary to store the counts of each number in the window.
2. Initialize a count variable to zero.
3. Iterate through the first window and update the counts in the dictionary.
4. If a count in the dictionary is 1, increment the count variable.
5. For the remaining windows, slide the window by one element to the right and update the counts in the dictionary accordingly.

6. If the count of the number that just left the window is 1, decrement the count variable.
7. If the count of the number that just entered the window is 1, increment the count variable.
8. Repeat steps 5-7 until you reach the end of the sequence.

Here's some Python code that implements this approach:



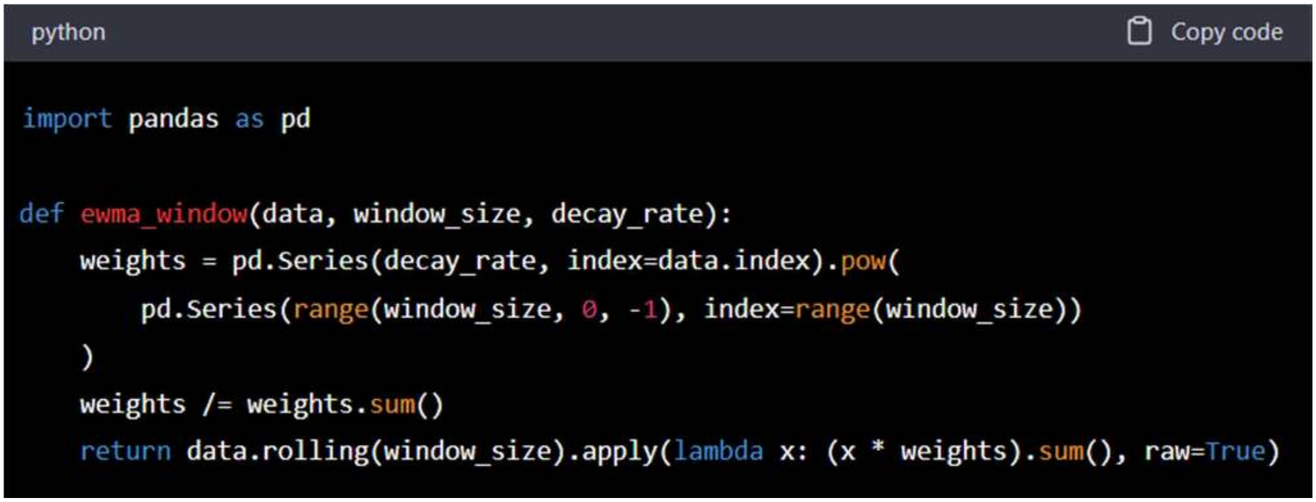
```python
def count_oneness(seq, window_size):
    counts = {}
    count = 0
    for i in range(window_size):
        num = seq[i]
        counts[num] = counts.get(num, 0) + 1
        if counts[num] == 1:
            count += 1
    for i in range(window_size, len(seq)):
        num_left = seq[i-window_size]
        num_enter = seq[i]
        counts[num_left] -= 1
        if counts[num_left] == 1:
            count -= 1
        counts[num_enter] = counts.get(num_enter, 0) + 1
        if counts[num_enter] == 1:
            count += 1
    return count
```

This function takes in a sequence seq and a window size window_size, and returns the number of times a number appears exactly once in a window of size window_size in the sequence. Note that this code assumes that all the elements in the sequence are integers. If the elements are not integers, you may need to modify the code accordingly.

Decaying Window

A decaying window is a common technique used in time-series analysis and signal processing to give more weight to recent observations while gradually reducing the importance of older observations. This can be useful when the underlying data generating process is changing over time, and more recent observations are more relevant for predicting future values.

Here's one way you could implement a decaying window in Python using an exponentially weighted moving average (EWMA):



```python
import pandas as pd

def ewma_window(data, window_size, decay_rate):
    weights = pd.Series(decay_rate, index=data.index).pow(
        pd.Series(range(window_size, 0, -1), index=range(window_size))
    )
    weights /= weights.sum()
    return data.rolling(window_size).apply(lambda x: (x * weights).sum(), raw=True)
```

This function takes in a Pandas Series data, a window size window_size, and a decay rate decay_rate. The decay rate determines how much weight is given to recent observations relative to older observations. A larger decay rate means that more weight is given to recent observations.

The function first creates a series of weights using the decay rate and the window size. The weights are calculated using the formula decay_rate^(window_size - i) where i is the index of the weight in the series. This gives more weight to recent observations and less weight to older observations.

Next, the function normalizes the weights so that they sum to one. This ensures that the weighted average is a proper average.

Finally, the function applies the rolling function to the data using the window size and a custom lambda function that calculates the weighted average of the window using the weights.

Note that this implementation uses Pandas' built-in rolling and apply functions, which are optimized for efficiency. If you're working with large datasets, this implementation should be quite fast. If you're working with smaller datasets or need more control over the implementation, you could implement a decaying window using a custom function that calculates the weighted average directly.

Real time Analytics Platform (RTAP) Applications

Real-time analytics platforms (RTAPs) are becoming increasingly popular as businesses strive to gain insights from streaming data and respond quickly to changing conditions. Here are some examples of RTAP applications:

1. **Fraud detection:** Financial institutions and e-commerce companies use RTAPs to detect fraud in real-time. By analyzing transactional data as it occurs, these companies can quickly identify and prevent fraudulent activity.
2. **Predictive maintenance:** RTAPs can be used to monitor the performance of machines and equipment in real-time. By analyzing data such as temperature, pressure, and

vibration, these platforms can predict when equipment is likely to fail and alert maintenance teams to take action.

3. **Supply chain optimization:** RTAPs can help companies optimize their supply chain by monitoring inventory levels, shipment tracking, and demand forecasting. By analyzing this data in real-time, companies can make better decisions about when to restock inventory, when to reroute shipments, and how to allocate resources.
4. **Customer experience management:** RTAPs can help companies monitor customer feedback in real-time, enabling them to respond quickly to complaints and improve the customer experience. By analyzing customer data from various sources, such as social media, email, and chat logs, companies can gain insights into customer behavior and preferences.
5. **Cybersecurity:** RTAPs can help companies detect and prevent cyberattacks in real-time. By analyzing network traffic, log files, and other data sources, these platforms can quickly identify suspicious activity and alert security teams to take action.

Overall, RTAPs can be applied in various industries and domains where real-time monitoring and analysis of data is critical to achieving business objectives. By providing insights into streaming data as it happens, RTAPs can help businesses make faster and more informed decisions.

Case Studies - Real Time Sentiment Analysis

Real-time sentiment analysis is a powerful tool for businesses that want to monitor and respond to customer feedback in real-time. Here are some case studies of companies that have successfully implemented real-time sentiment analysis:

1. **Airbnb:** The popular home-sharing platform uses real-time sentiment analysis to monitor customer feedback and respond to complaints. Airbnb's customer service team uses the platform to monitor social media and review sites for mentions of the brand, and to track sentiment over time. By analyzing this data in real-time, Airbnb can quickly respond to complaints and improve the customer experience.
2. **Coca-Cola:** Coca-Cola uses real-time sentiment analysis to monitor social media for mentions of the brand and to track sentiment over time. The company's marketing team uses this data to identify trends and to create more targeted marketing campaigns. By analyzing real-time sentiment data, Coca-Cola can quickly respond to changes in consumer sentiment and adjust its marketing strategy accordingly.
3. **Ford:** Ford uses real-time sentiment analysis to monitor customer feedback on social media and review sites. The company's customer service team uses this data to identify issues and to respond to complaints in real-time. By analyzing real-time sentiment data, Ford can quickly identify and address customer concerns, improving the overall customer experience.
4. **Hootsuite:** Social media management platform Hootsuite uses real-time sentiment analysis to help businesses monitor and respond to customer feedback. Hootsuite's sentiment analysis tool allows businesses to monitor sentiment across social media

channels, track sentiment over time, and identify trends. By analyzing real-time sentiment data, businesses can quickly respond to customer feedback and improve the overall customer experience.

5. **Twitter:** Twitter uses real-time sentiment analysis to identify trending topics and to monitor sentiment across the platform. The company's sentiment analysis tool allows users to track sentiment across various topics and to identify emerging trends. By analyzing real-time sentiment data, Twitter can quickly identify issues and respond to changes in user sentiment.

Overall, real-time sentiment analysis is a powerful tool for businesses that want to monitor and respond to customer feedback in real-time. By analyzing real-time sentiment data, businesses can quickly identify issues and respond to changes in customer sentiment, improving the overall customer experience.

Case Studies - Stock Market Predictions

Predicting stock market performance is a challenging task, but there have been several successful case studies of companies using machine learning and artificial intelligence to make accurate predictions. Here are some examples of successful stock market prediction case studies:

1. **Kavout:** Kavout is a Seattle-based fintech company that uses artificial intelligence and machine learning to predict stock performance. The company's system uses a combination of fundamental and technical analysis to generate buy and sell recommendations for individual stocks. Kavout's AI algorithms have outperformed traditional investment strategies and consistently outperformed the S&P 500 index.
2. **Sentient Technologies:** Sentient Technologies is a San Francisco-based AI startup that uses deep learning to predict stock market performance. The company's system uses a combination of natural language processing, image recognition, and genetic algorithms to analyze market data and generate investment strategies. Sentient's AI algorithms have consistently outperformed the S&P 500 index and other traditional investment strategies.
3. **Quantiacs:** Quantiacs is a California-based investment firm that uses machine learning to develop trading algorithms. The company's system uses machine learning algorithms to analyze market data and generate trading strategies. Quantiacs' trading algorithms have consistently outperformed traditional investment strategies and have delivered returns that are significantly higher than the S&P 500 index.
4. **Kensho Technologies:** Kensho Technologies is a Massachusetts-based fintech company that uses artificial intelligence to predict stock market performance. The company's system uses natural language processing and machine learning algorithms to analyze news articles, social media feeds, and other data sources to identify patterns and generate investment recommendations. Kensho's AI algorithms have consistently outperformed the S&P 500 index and other traditional investment strategies.

5. **AlphaSense:** AlphaSense is a New York-based fintech company that uses natural language processing and machine learning to analyze financial data. The company's system uses machine learning algorithms to identify patterns in financial data and generate investment recommendations. AlphaSense's AI algorithms have consistently outperformed traditional investment strategies and have delivered returns that are significantly higher than the S&P 500 index.

Overall, these case studies demonstrate the potential of machine learning and artificial intelligence to make accurate predictions in the stock market. By analyzing large volumes of data and identifying patterns, these systems can generate investment strategies that outperform traditional methods. However, it is important to note that the stock market is inherently unpredictable, and past performance is not necessarily indicative of future results.