**Project Report: Craft Beer Analysis using Hive and Spark**

**Introduction**

The project, "Craft Beer Analytics with Hive and Spark," aims to demonstrate the students' ability to use sophisticated big data tools and techniques for in-depth analysis. The main goal is to use Apache Hive and Apache Spark to process, analyse, and visualise large amounts of data pertaining to the production of craft beer.

The dataset includes important factors that are necessary to comprehend and improve the craft brewing process. Batch\_ID, Brew\_Date, Beer\_Style, SKU, Location, Fermentation\_Time, Temperature, pH\_Level, Gravity, and Alcohol\_Content are some of these parameters. Every entry in the dataset represents a distinct batch of craft beer, offering insightful information about the qualities of the finished product and the brewing process.

I will delve into details of the craft beer industry by utilising the capabilities of Apache Hive and Apache Spark. I will look for trends, recognise patterns, and extract useful information from the dataset. The data's diverse nature, which includes variables like Beer Style, Location, and Fermentation Time, will allow students to conduct in-depth analysis and draw insightful conclusions about the variables influencing the production of craft beer.

**Background**

Breweries face a distinct set of opportunities and challenges in the constantly changing craft beer industry. A multitude of breweries have emerged as a result of the renewed interest in craft beer, each of which offers a wide variety of beer flavours and styles. In this ever-changing environment, the importance of insightful analysis and data-driven decision-making increases.

With its focus on quality, variety of flavours, and frequently smaller-scale production, craft beer offers a rich dataset with subtleties that go beyond traditional brewing. Breweries have to strike a delicate balance between operational effectiveness and creative expression. The integration of big data analytics becomes strategically imperative in order to navigate this complexity and capitalise on emerging trends.

**Data Acquisition and Cleaning**

This dataset, which covers the period from January 2020 to January 2024, offers an extensive collection of information from a craft beer brewery. It provides a comprehensive understanding of the brewing process and its effects on the market by encapsulating a rich blend of brewing parameters, sales data, and quality assessments.

**Data Format and Structure**

The dataset comes in a CSV file with a tabular structure that makes it simple to integrate with a variety of data analysis programmes.

It consists of more than 10 million records, each of which is a distinct batch with an extensive feature set.

Below are the top 10 rows of the dataset from hive and spark.

A screen shot of a computer

Description automatically generated

**Objectives**

* Utilize Apache Hive and Apache Spark for processing and analyzing craft beer data.
* Identify trends, patterns, and critical factors influencing craft beer production.
* Extract meaningful information from the diverse dataset, considering variables like Beer Style, Location, and Fermentation Time.
* Conduct in-depth analysis to draw insightful conclusions about the variables affecting craft beer production.

**Methodology**

* Data Processing with Apache Hive
  + Data Ingestion: Import the craft beer dataset into Apache Hive for efficient storage and retrieval.
  + Data Exploration:Conduct exploratory data analysis to understand the structure and characteristics of the dataset.
  + Data Cleaning:Address missing or inconsistent data to ensure data quality.
  + Data Transformation:Create tables in Hive to facilitate seamless querying.
* Analysis with Apache Spark
  + Data Transformation: Transform Hive data for compatibility with Apache Spark.
  + Feature Engineering: Extract relevant features and create new variables for analysis.
  + Statistical Analysis: Utilize Spark's capabilities for statistical analysis, including descriptive statistics and correlation.
* Visualization
  + Data Visualization: Utilize visualization tools (e.g., Matplotlib, Seaborn) to represent trends and patterns.

**Challenges and Solutions:**

* Challenge 1: Diverse Dataset Size
  + Challenge: The craft beer dataset is extensive, making it challenging to manage efficiently.
* Solution:
  + Data Partitioning: Split the dataset into manageable partitions for parallel processing in Hive.
* Challenge 2: Complex Brewing Process Information
  + Challenge: The dataset contains intricate information about the brewing process, making transformations complex.
* Solution:
  + Divide and Conquer: Break down complex transformations into smaller, manageable phases for clarity and efficiency.
  + Hive Functions: Leverage built-in Hive functions for common transformations to streamline the process.
* Challenge 3: Query Performance Issues
  + Challenge: Performance issues may arise during querying and data retrieval.
* Solution:
  + Hive Setup Optimization: Fine-tune Hive configurations to better suit the dataset and improve overall performance.
  + Query Optimization: Optimize queries by analyzing execution plans and making necessary adjustments.
* Challenge 4: Variability in Data Distribution
  + Challenge: Skewness in data distribution may affect the performance of certain operations.
* Solution:
  + Data Distribution Modification: Modify the distribution of data, if possible, to achieve a more balanced workload.
* Challenge 5: Limited Insights with Basic Analytics
  + Challenge: Basic analytics may not provide sufficient insights; advanced analytics capabilities are needed.
* Solution:
  + Integration with Apache Spark: Integrate Apache Spark for advanced analytics tasks that go beyond Hive's capabilities.