**Abschlussprojekt - Big Data Infrastruktur**

**Thema: Analyse von Lebensmittelpreisen in österreichischen Supermärkten**

**Teammitglieder**

* Ambar Irfan
* Julia Brandstätter
* Pinar Yalcin

**Geplante Datenquellen**

1. **Web Scraping von Supermarkt-Websites** – Extrahieren von Preisen für Grundnahrungsmittel (z. B. Milch, Brot, Eier) von Online-Shops wie Billa, Spar oder Hofer.
2. **Open Data Österreich / Statistik Austria** – Nutzung öffentlich zugänglicher Preisstatistiken für den Lebensmittelbereich.

**Geplante Datenspeicherung**

* **Datenbank:** MongoDB (NoSQL)
  + Speicherung der extrahierten Supermarktpreise.
  + Effiziente Abfragen nach Produkt, Supermarkt und Preisentwicklung.
* **Cloud-Speicher:** Google Drive / AWS S3
  + Speicherung der Rohdatensätze für Sicherungszwecke.

**Geplante Vorgehensweise**

1. **Datenerfassung**
   * Implementierung eines **Web Scraping Scripts** mit **Selenium / BeautifulSoup**, um Preise von Supermarkt-Webseiten zu extrahieren.
   * Herunterladen und Vorverarbeitung von Open Data-Preisstatistiken aus Statistik Austria.
2. **Datenverarbeitung**
   * Speicherung der Daten in MongoDB zur strukturierten Abfrage.
   * Verwendung von Pandas zur Datenbereinigung und Formatierung.
   * Implementierung eines einfachen **MapReduce**-Algorithmus zur Analyse der Preisentwicklung.
3. **Datenanalyse**
   * Vergleich der Lebensmittelpreise zwischen verschiedenen Supermärkten.
   * Identifikation von Preisänderungen über die Zeit.
   * Durchschnittspreisberechnungen für verschiedene Produktkategorien.
4. **Visualisierung & Berichterstellung**
   * Nutzung von Matplotlib/Plotly für Preisdiagramme und Visualisierungen.
   * Erstellung eines Jupyter Notebooks zur Dokumentation der Analyse.

**Erwartete Ergebnisse**

* **Vergleich von Lebensmittelpreisen zwischen verschiedenen Supermärkten**
* **Erkenntnisse zur Preisentwicklung über Zeiträume**
* **Grafische Darstellung von Preisverläufen und Trends**
* **Vollständig dokumentiertes Jupyter Notebook**
* **Abschlusspräsentation mit zentralen Erkenntnissen**

**Verwendete Technologien & Tools**

* **Python, Pandas, Selenium, BeautifulSoup** (Datenextraktion & Analyse)
* **MongoDB (NoSQL)** (Datenspeicherung)
* **Git & GitHub** (Versionskontrolle & Zusammenarbeit)
* **Docker** (Infrastruktureinrichtung & Bereitstellung)
* **Matplotlib, Plotly** (Datenvisualisierung)

Dieses Dokument beschreibt den grundlegenden Plan für unser **Big Data Infrastruktur**-Projekt. Der Schwerpunkt liegt auf der Einrichtung der Infrastruktur für die Erfassung, Speicherung und Verarbeitung von Lebensmittelpreisen, während ein strukturierter Workflow mit Git und Docker eingehalten wird.