UNIVERSIDAD
D
CÓRDOBA

Instituto de Estudios de Posgrado
Universidad de Córdoba

Máster Universitario en Inteligencia Computacional e Internet de las Cosas

Planning your pipeline

Análisis, Diseño y Procesamiento de Datos Aplicados a las Ciencias y a las Tecnologías

Autora:
Alba Márquez-Rodríguez

Profesores:
Gonzalo Cerruela García
Dormingo Ortiz Boyer
Juan A. Romero del Castillo

Córdoba, Enero 2024

Índice

1. The Elements of Data Activity: Planning Your Pipeline

1.1. Task

You are a data engineer and are assigned to build a pipeline for a selected scenario.

1. Choose one of the scenarios that is listed in the Scenarios section.
2. Use the worksheet to document your ideas about the general requirements and questions to ask for each of the five Vs of data as they relate to your scenario:
   - Value – What insights can be pulled from the data?
   - Veracity – How accurate, precise, and trusted is the data?
   - Variety – What types and formats? How many different sources of data?
   - Velocity – What is the frequency of new data being generated and ingested?
   - Volume – How big is the dataset? How much new data is generated?

     The Worksheet instructions section includes prompting questions to help guide your thinking.
3. As directed by your instructor, discuss or present your results to the class.

1.2. Scenario

1.2.1. Gaming analytics

Your startup company has launched a game that quickly became very popular. The company needs to update the in-game experience to stay competitive. It wants to update designs and rapidly improve gameplay. Currently, the company receives feedback through a feedback button that is embedded in the game navigation and lets users send comments. Gameplay data is sent and received as JSON messages.

Desired outcome: Developers can analyze gaming behaviors in near real time. They can rapidly make design updates that are based on analysis of gameplay data and feedback that was submitted through the feedback button.

1.3. Worksheet instructions

1.3.1. Selected scenario

Indicate which scenario you are evaluating (e-bike rental, gaming analytics, or fraud detection).

1.3.2. Potential data sources

List data sources that might be valuable to meet the desired outcome. For each source, include what you think the format would be. Add as many data sources as you can think of that might be useful, even if you can't fill in details about the data source.

- Think about what types of public datasets might be useful.
- Think about what types of internal or company-owned data might be available.
- Think about any event or time-series data that might be relevant.

1.3.3. Value, veracity, and variety considerations

Describe what you need to discover about the sources that you listed to determine if they would be valuable for your use case. Describe the tasks that you think would be needed to get value from the data. Here are some questions to think about:

- What type of analytics and how much processing might be needed to derive insights?
- What type of queries could be made for analysis and visualization?

- What kind of privacy laws might affect data collection and retention across different geographies and industries?
- What kind of security considerations must be addressed? Does the source contain personally identifiable information (PII)?
- What types of veracity challenges might need to be addressed?
- How quickly would the data become stale?
- What details do you know about the data (for example, fields and field types)?
- What type of cleaning and transformations might be needed to use this data?
- How will this data source be combined with other data sources?

1.3.4. Velocity and volume considerations

For the data sources that you identified, describe the general requirements for volume and velocity expectations in the context of your selected scenario. Here are some questions to think about:

- What would be the general volume and velocity of incoming data? How might data be transferred into your system (for example, FTP or a connected device)?
- Would the volume of data that is ingested into the pipeline be at steady level, or would it be spiky? Are the patterns predictable?
- Would data be ingested and processed at a regular cadence (for example, nightly or weekly)?
- Does the data need to be processed and analyzed in near real time to be valuable?

2. Worksheet

Selected Scenario: Gaming Analytics
Potential Data Sources
1. In-game Feedback Button: Structured (JSON messages) **2. Gameplay Data - Format**: Semi-structured (JSON messages) **3. Player Profiles:** Structured (relational database) **4. In-app Purchases:** Structured (transaction records) **5. Social Media Mentions:** Unstructured (text data from various platforms, social media platforms APIs)
Value, Veracity, and Variety Considerations
1. In-game Feedback Button: - **Value**: Insights on player preferences, suggestions, and possible improvements. - **Veracity**: Depends on the authenticity of user-provided feedback. The longer the account has been active, the greater the probability of veracity. The less time the account has been active and the more posts it makes, the less likely it is to be true. - **Variety**: Written comments, ratings, and potentially attached screenshots and videos. **2. Gameplay Data:** - **Value**: Real-time analysis of player actions, behaviors, and interactions within the game. - **Veracity**: High, as it's generated by the game system itself. - **Variety**: JSON messages containing various gameplay events (e.g., player movements, in-game purchases, interactions...). **3. Player Profiles:** - **Value**: Understanding player demographics, preferences, and history. - **Veracity**: Reliable if the profiles are properly maintained and similar to the In-game Feedback Button. The longer the account has been active and the most activity it has, the greater the probability of veracity. - **Variety**: Structured data including player ID, player data, achievements, playtime, preferences... **4. In-app Purchases:** - **Value**: Insights into popular items, purchase patterns, and revenue generation. - **Veracity**: High, as it involves financial transactions from players. - **Variety**: Structured data with transaction details, item IDs, and purchase amounts. **5. Social Media Mentions:** - **Value**: Perception of people from and not from the game, trends, and potential issues, suggestions and improvements discussed by players. - **Veracity**: Subject to the reliability of external sources; may include misinformation. It could work as In-game Feedback, the longer the account has been active, the greater the probability of veracity. - **Variety**: Unstructured text data from various platforms, including tweets, posts, and comments. It may vary depending on the API, but could be preprocessed to make it more consistent.
Velocity and Volume Considerations
1. In-game Feedback Button and Gameplay Data: - **Volume**: High, especially during peak gaming times or after updates. - **Velocity**: Near real-time processing required for rapid response to player feedback. **2. Player Profiles and In-app Purchases:** - **Volume**: Moderate, as these are generated less frequently than gameplay data. - **Velocity**: Can be processed at regular intervals (e.g., nightly) due to less time-sensitive nature. **3. Social Media Mentions:** - **Volume**: Variable, depending on the game's popularity and current events. - **Velocity**: Near real-time processing would be beneficial for timely responses to emerging issues.