BI Data Engineer Case Study

# Úkol 1 (MS Azure)

V prostředí MS Azure pomocí poskytnutého endpointu pošlete z Vámi vytvořeného notebooku do Azure Event Hubu 1000 event (zpráv) ve formátu json. Každý event by měl obsahovat čas vytvoření, alespoň jednu kategorickou proměnnou (například ID hry) a alespoň jednu spojitou proměnnou (například výše sázky).

Následně Vámi vytvořená data opět načtěte a vytvořte libovolnou filtraci a napočtěte libovolné metriky. Výsledky uložte do csv souboru.

HINT 1: Využijte jedno pole jsonu pro filtraci mezi Vašimi jednotlivými pokusy. Podle tohoto pole pak načtete pouze relevantní data pro zpracování.

HINT 2: Pokud se vám nepodaří vygenerovat data a zaslat je do Azure Event Hubu, kontaktujte nás pro možnost splnění zbytku úlohy.

Výsledkem by měl být Python skript, který posílá eventy, filtruje je a ukládá do csv souboru.

***Endpoint detaily:***

Event hub namespace connection string:

Endpoint=sb://szbidecs.servicebus.windows.net/;SharedAccessKeyName=GenericAccessKey;SharedAccessKey=DYuMcEowO2kyrKOkUNdWTWzCDqgUA5tCs+AEhGH6qo4=

Název event hubu:

szcasestudy{prijmeni}, kde {prijmeni} nahraďte Vaším příjmením (malými písmeny, bez diakritiky), např tedy.: szcasestudynovakova

# Úkol 2

Váš byznysový partner chce získávat real-time metriky o součtech sázek per produkt za posledních 5 minut. Nasimulujte tok dat (eventy z úkolu 2) tak, aby se do Azure Event Hubu posílaly rovnoměrně během více než 15 minut. Zároveň vytvořte nástroj, který poskytne každých 20 sekund update o součtech sázek pro každou hru zvlášť. Výsledky posílejte zpět do poskytnutého Azure Event Hubu.

HINT: Stejně jako v úkolu 1 využijte pole výsledného jsonu pro rozlišení mezi vstupními a výstupními eventy.

Výsledkem by měl být Python skript, který zobrazuje výsledky.

# Úkol 3

Navrhněte end-to-end samoučící pipeline, která bude měsíčně analyzovat data a vybírat POS s nejvyšším potenciálem zvýšení měsíčních příjmů. Výstupní data budou ukládána a zpřístupňována automaticky každý měsíc byznysovému vlastníkovi. Technická implementace tohoto řešení bude probíhat v prostředí MS Azure/Databricks.

Výchozí stav: Sazka on-premise databáze obsahuje tabulku s detailními informacemi o každém Point of Sales (POS) v retailové síti. Tato tabulka zahrnuje ID každého POS, jeho GPS souřadnice a informace o denních obratech pro každý existující produkt od Sazky.

Výsledkem může být např. textový popis/diagram dle Vašeho uvážení.