Zápočtový test

Na vypracování testu máte 120 minut. Očekávanými a hodnocenými výstupy jsou:

1. Textový soubor s názvem *váš_login*.txt (např. pascepet.txt) obsahující zdrojový kód příkazů, kterými jste provedli zadání. Pokud máte zároveň něco zjistit nebo vypsat, zaznamenejte to přímo do souboru, ideálně jako komentář. **Tento soubor na konci své práce zašlete mailem na adresu jan.hucin@profinit.eu a podepište se svým celým jménem.**
2. Existence a vlastnosti souborů, adresářů, tabulek, které jste zkopírovali či vytvořili při plnění zadání. Ty zhodnotíme přímo na clusteru.

V jednotlivých oblastech testu se hodnotí každý úkol nebo jeho část. Pokud si s nějakou částí zadání nebudete vědět rady, můžete ji přeskočit nebo zadání splnit bez této části, počet bodů se pak přiměřeně sníží.

HDFS operace (3 body)

- Na svém uživatelském adresáři HDFS (/user/*váš_login*) nastavte pro veřejnost právo read a execute.
- Ve svém uživatelském adresáři na HDFS (/user/*váš_login*) založte podadresář **kingbase**.
- Zkopírujte z lokálního filesystemu na metacentru z podadresáře /home/pascepet/fel_bigdata/data/kingbase všechny soubory se jménem začínajícím **KingBase2016-03-E** na HDFS do podadresáře, který jste založili v předchozím kroku.

V dalších třech částech budete pracovat se stejnými daty. Na HDFS je adresář **/user/pascepet/data/kingbase** obsahující soubor se záznamy o odehraných hrách v šachových turnajích. Soubor je standardní textový s oddělovači (znak '~'), obsahuje hlavičky (názvy) sloupců. Sloupce obsahují postupně jména hráčů s bílými a černými figurami, ratingy obou hráčů, datum hry, výsledek, řetězec s tagy ke hře, řetězec se záznamem tahů ve hře, body pro bílého a počet tahů.

Práce s Hive (9 bodů)

Zkontrolujte, že máte založenou databázi Hive (název = váš login), a pokud ne, založte ji. Dále pracujte se svou databází.

- Vytvořte externí tabulku Hive *games_ext* založenou na datech z výše uvedeného adresáře (v rámečku). Typy polí zvolte vhodně podle svého uvážení.
- Vytvořte managed (interní) tabulku Hive *games* s formátem Parquet a kompresí GZIP. Oproti externí tabulce bude mít navíc sloupec na počet bodů pro černého. Do této tabulky přeneste data z externí tabulky, ale jen ty řádky, kde mají oba hráči rating aspoň 2500 a počet tahů ve hře byl aspoň 30. Počet bodů pro černého se určí ze vztahu počet bodů pro bílého + počet bodů pro černého = 1.
- Po úspěšném přenosu dat do managed tabulky zrušte tabulku *games_ext*.

Z výsledné tabulky *games* zjistěte pomocí SQL dotazu:

- Jaká je četnost jednotlivých výsledků?
- Kterých pět hráčů má v záznamech nejvíce her odehraných černými figurami?

Spark RDD (9 bodů)

Spark spouštějte v konfiguraci *--num-executors 2 --executor-memory 4G --packages com.databricks:spark-csv_2.10:1.5.0 --conf spark.ui.port=1<ddmm>*, kde <ddmm> je váš den a měsíc narození, napr. spark.ui.port=10811

- Načtěte obsah výše uvedeného adresáře (v rámečku) do RDD.
- Zjistěte pět nejčastějších kombinací prvních dvou tahů (první dva tahy jsou např. 1.d4 Nf6 nebo 1.e4 c5)
- Kolik her skončilo matem (záznam tahů ve hře obsahuje znak '#')?
- Vezmeme-li jen hry, kde oba hráči měli rating nad 2600, kolik tahů dohromady v nich bylo provedeno jezdcem (zápis tahu obsahuje písmeno N)? Jednotlivé tahy jsou v záznamu tahů odděleny mezerami.

Spark SQL (9 bodů)

- Načtěte obsah výše uvedeného adresáře (v rámečku) do DataFrame.
- Kolik her celkem (jako bílý nebo jako černý) odehrál hráč jménem „Nakamura, Hikaru“ v roce 2015?
- Jaký byl průměrný bodový zisk hráčů s bílými figurami za hry, kde se ratingy obou hráčů lišily nejvýše o 50 bodů?
- Jaký byl největší počet tahů v jedné hře? Kteří hráči ji hráli a jakým výsledkem skončila?