Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Evidenčné číslo

Matúš Cimerman

Analýza prúdu údajov

Bakalárska práca

Vedúci práce: Ing. Jakub Ševcech
máj 2015

Abstract

Nowadays we can see Big Data processing and analysis in many domains. As a amounts of data growing, more people are focusing on this problem. The most affected domains are social media websites like Facebook or Twitter. A data from such a sources are streaming in huge amounts and changing in real-time, called data streams. We want to process and analyze data streams in real-time to provide users personalized and valuable outputs. The most common approach to handle data streams is map-reduce paradigm, e.g. batch data processing. Proposed methods are not meeting our requirement to process data streams in real-time. To achieve these requirements, we need use different approach called data stream processing which is built on Lambda Architecture.
Processing and analysis of big data streams is a complex task, because we need to provide low-latency, scalable and fault-tolerant solution. In our project, we analyze existing solutions and frameworks to analyze data streams. We provide verification of its characteristics in different kind of tasks. Accordingly to this, we propose a application for processing and analyzing big data streams (e.g. Twitter data stream), which allows users to get valuable outputs changing in real-time.

ToDo: upravit anglicku podla slovenskej. V slovenskej su male upravy...-

az po revizii slovenskej

Abstrakt

Dnes sa stretávame so spracovaním a analýzou veľkého objemu dát v mnohých oblastiach. S narastajúcim objemom dát rastie aj záujem o tento problem spracovania veľkého objemu dát. Toto najviac postihuje oblasť sociálnych medií, napríklad Facebook, či Twitter. Dáta z takýchto zdrojov prúdia v obrovkých množstvách, pričom dáta sa v čase menia, takéto prúdiace dáta nazývame jednoducho prúd dát. Prúdy dát chceme spracovať a analyzovat v reálnom čase, aby sme mohli ponúknuť používateľom personalizovaný a hodnotný výstup. Najčastejší prístup spracovania prúdu dát je map-reduce paradigma, napríklad dávkové spracovanie. Táto metóda, ale nespĺňa naše požiadavky na spracovanie dát v reálnom čase. Na dosiahnutie našej požiadavky potrebujeme iný prístup, ktorý nazývame spracovanie prúdu dát. Metóda takéhoto prístupu je postavená na Lambda architektúre. Spracovanie a analýza veľkých prúdov dát je komplexný problém, pretože prúd musí byť spracovaný s nízkou odozvou, riešenie musí byť odolné proti chybám a škálovateľné. V našej práci budeme analyzovať existujúce riešenia and rámce na analýzu prúdu dát. Poskytujeme overenie charakteristík týchto riešení v roznych problémoch, ktoré si vyžadujú spracovanie prúdu dát v reálnom čase. Na základe tohto poskytujeme aplikáciu, ktorá spracuje a analyzuje veľký prúd dát (napríklad prúd dát Twittru), ktorá umožní používateľom získať hodnotný výstup, ktorý sa mení v reálnom čase.

ToDo: zre-v-i-dovat sloven-sku verziu

Poďakovanie

Thank all of you!

Matúš Cimerman

Obsah

1. Úvod

ToDo: napisat uvod

2. Metódy spracovania prúdu údajov

3. Existujúce riešenia pre spracovanie a anlýzu prúdu údajov

4. Metóda pre spracovanie prúdu údajov v reálnom čase

V predchádzajúcej kapitole sme sa venovali existujúcim riešeniam na spracovanie veľkého objemu dát. Niektoré sa zameriavajú na dávkové spracovanie veľkých dát, iné na spracovanie "rýchlych dát" v reálnom čase. Navrhujeme metódu na analýzu prúdu údajov. Dáta prúdia zo zdroja v reálnom čase. Naša metóda spracuje prúd údajov v reálnom čase a poskytytuje používateľovi hodnotný výstup na základe dopytu. Metódu navrhujeme pre doménu sociálnych sietí, konkrétne Twitter. Cieľom je vizualizovať správy na mape, ktoré majú súvislosť s dopytom používateľa.

5. Implementácia navrhnutej metódy

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6. Vyhodnotenie a experimenty

7. Zhodnotenie a budúca práca

7.1 Zhodnotenie a záver

7.2 Budúca práca

A. Plán na letný semester

ToDo: Plan na druhy semester!

B. Druhá príloha

Bibliography