Apache Hadoop

Apache Hadoop je Big data framework, ktorý umožňuje vytvárať aplikácie, ktoré spracovávajú obrovské množstvo dát nad klastrom bežne dostupného komoditného harvéru. Základom je distribuovaný súborový systém a nástroje na paralelné spracovávanie dát. Takáto architektúra poskytuje agregovanú kapacitu úložiska a agregovaný výpočtový výkon. Veľké úlohy sú rozdelené na množstvo menších úloh, ktoré sú vykonané na klastri s lacným komoditným hardvérom. Apache Hadoop je open source riešenie, takže na jeho prevádzku nie sú nutné žiadne komerčné licencie. Existuje však aj viacero distribúcií, ktoré prinášajú pridanú hodnotu vo forme technickej podpory alebo vlastných komerčných nástrojov.

Základné prínosy Hadoop ako platformy:

Flexibilita:

* dáta môžu byť ukladané bez potreby vopred poznať štrukúru
* dátový model je aplikovaný na základe konkrétnych otázok
* vysoká horizontálna škálovateľnosť. Je možné začať malým „pieskoviskom“, ktoré môže časom narásť až do plnohodnotného dátového jazera.

Efektívnosť:

* 100% dát je dostupných na analytiku s maximálnou granularitou
* možnosť analyzovať štruktúrované aj neštruktúrované dáta v rôznych prípadoch rôznymi spôsobmi

Cenová efektivita:

* 100% open source
* možnosť prevádzky na lacnom komoditnom HW
* nové dáta môžu byť prístupné na analytiku ihneď bez potreby vývoja ETL procesov

### Infraštruktúra

Hadoop architektúra umožňuje presunúť výpočtovú silu bližšie k zdrojovým dátam. Výpočtová časť Hadoop klastra môže byť nainštalovaná na bežnom komoditnom harvéri, ktorý poskytuje lacnú výpočtovú silu a lacné úložisko v porovnaní s high-end servermi. Hadoop je veľmi ľahko horizontálne škálovateľný, rozšírenie kapacity klastra je možné robiť bez výpadku. Hadoop platforma môže byť prevádzkovaná on-premises v privátnych dátových centrách alebo vo verejných cloudových platformách.

### Integrácia

Hadoop platforma disponuje nástrojmi, vďaka ktorým je možné nasmerovať dáta z akýchkoľvek zdrojových systémov na Hadoop platformu, či už dávkovo alebo v reálnom čase kde sú dáta ihneď dostupné dátovým analytikom.

### **Pokročilá dátová analytika**

Analytické nástroje v spojení so schopnosťou Hadoop platfomy spracovávať veľké množstvo dát si poradia s akoukoľvek analytickou úlohou. Hadoop umožňuje spracovávať oveľa väčšie objemy dát ako bežné analytické platformy, čím sa otvárajú nové možnosti pri plnení analytických úloh.

### Vizualizácia a reporting

Hadoop je možné použiť ako zdroj dát pre bežne používané vizualizačné nástroje. Väčšina dostupných nástrojov obsahuje natívnu podporu Hadoop platformy. Pri našich projektoch sa osvedčili široko používané nástroje ako Tableau, Qlik Sense či Power BI.

## Naše služby

Našim zákazníkom ponúkame služby, ktoré pokývajú všetky oblasti od vybudovania Hadoop platformy na mieru až po sofistikovanú dátovú analytiku.

### Implementácia Hadoop

* služby implementácie Hadoop v prostredí u zákazníka resp. do cloudového prostredia (Hadoop as a service)

### Integrácia dátových zdrojov

* nasmerovanie dát zo zdrojových systémov na Hadoop platformu
* **nastavenie zálohovania**

### Dátová analytika

* deskriptívna a prediktívna analytika podľa požiadaviek zákazníka
* vizualizácia dát a reporting

### Dátová politika

* vytvorenie dátového katalógu
* definovanie procesov zabezpečeného prístupu k dátam na základe používateľských rolí

### Podpora prevádzky

* administrácia platformy
* zaškolenie interného tímu administrátorov a dátových analytikov
* zaškolenie interného analytického tímu v oblasti
* outsourcing analytických konzultantov

## Oblasti použitia

## Príklad použitia

Pre nášho zákazníka Orange Slovensko sme implementovali analytický datamart „Modelling data cache“ na Hadoop platforme, ktorý slúži ako zdroj údajov pre prediktívne modelovanie. Datamart je široká tabuľka, ktorá obsahuje viac ako 350 behaviorálnych atribútov o každom zákazníkovi. Súčasťou dodávky bol aj prediktívny model na predikciu odchodovosti zákazníkov a framework na tvorbu prediktívnych modelov. Zákazník má ambíciu ďalší vývoj robiť vo vlastnej réžii a tak bolo odovzdanie projektu ukončené sériou workshopov pre užívateľov dodanej platformy.

**Použité technológie:** Hadoop HDFS, HUE, Apache Hive, Apache Impala, Apache Spark, Python, Scikit-learn

Hadoop prináša možnosť skladovať a analyzovať štruktúrované aj neštruktúrované dáta. Tým je možné vyťažiť z dát maximum informačnej hodnoty a to všetko za zlomok ceny komerčných DWH riešení. Najbežnejšie oblasti použitia Hadoop platformy sú:

* pokročilá zákaznícka analytika
* vyhodnocovanie prevádzkových dát z výroby
* prediktívna údržba
* aktívny archív
* zber a vyhodnotenie dát z IOT zariadení
* odľahčenie existujúcich DWH platforiem
* podporná platforma pre transakčné systémy