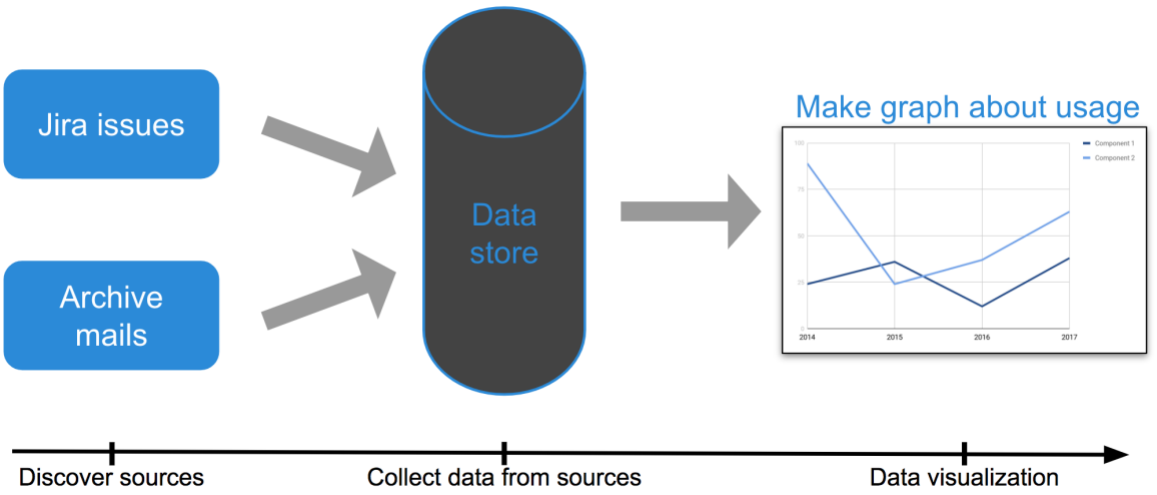
# Beszámoló kutatómunkáról

## CDH usage pattern analysis

A CDH usage pattern analysis című téma különböző felhasználói mintáknak a vizsgálatával foglalkozik. A tematikán belül két fő témakört fogalmaztunk meg, melyekkel a félévek során foglalkozni szeretnénk. Az első azoknak a CDH (a Cloudera Big Data platformja) komponenseknek a megtalálásáról szól, amelyeket gyakran kombinálnak egymással, gyakran használják őket együtt. A másik pedig a különböző CDH komponensek népszerűségének a változását mutatja meg az idő során. A két altémakör eredményeiből következtetéseket vonhatunk le számos kérdésben. Például, ha kiderítjük, hogy mely komponenseket használnak együtt másokkal, akkor feltételezéseket tehetünk azokról a komponensekről, amelyek a következő kiadásnál kivonhatóak a forgalomból, vagy azokról a komponensekről is kialakulhat egy képünk, melyek alapvető fontosságúak a Cloudera ügyfeleinek körében.

A félév során a témakör második részével foglalkoztam. Ehhez a félév elején célokat fogalmaztunk meg, mint a CDH környezetének és komponenseinek megismerése és azoknak az adatforrásoknak a megtalálása, melyekből használható információkat gyűjthetünk a népszerűségi értékek meghatározásához. Majd a félév utolsó pontjaként maguknak az adatoknak a gyűjtését tűztük ki, ezekből a kiválasztott adatforrásokból, struktúrált, azaz később is könnyen kezelhető módon. A félév eredményének pedig ezeknek az adatforrásokból származó adatoknak a vizualizációját tűztük ki.
A féléves munkám folyamatát e kép szemlélteti.

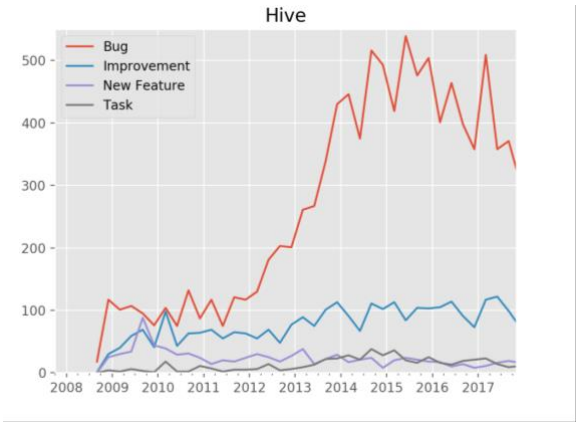
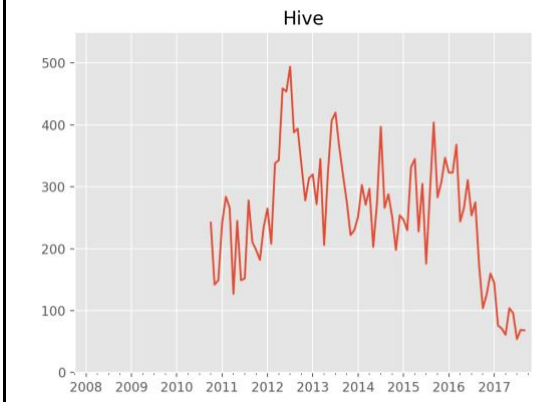



A Jira egy hiba bejelentő és probléma követő szoftver. Az Apache projektekhez használt rendszer sok információt árul el a komponensek használatáról, így a népszerűségükről is, ezért választottuk adatforrásnak. A Jira rendszer fejlesztői REST API-t biztosítanak a tikettek lekérdezéséhez, mely JSON formátumban válaszol a kérésekre. A lekérdezés JQL nyelven történik. A Python Jira modulját használtam a tikettek letöltésére. A modul a REST API-s hozzáférést használja a lekérdezésekhez. Négy projektet vizsgáltam a Pig-et, amit Big Data elemzésre fejlesztenek, a Hive-ot, amivel SQL adatbázisként tekinthetünk Hadoop-ban tárolt adatainkra, az Impala-át, amivel valós időben kérdezhetünk le az adattárházakból, és az Avro-t, ami adat szerializációs komponens. A komponenseket

az első tikett feladásától elemeztem. A projektek letöltött információit Excel táblához hasonló formátumban tároltam, melyhez a Python Pandas modult használtam fel.

A félév második adatforrása a levelezési listák archívuma volt. Minden Apache projektnek van levelező listája, ami adott időszaktól függően szokott forgalmas lenni. Az archivált listák honlapjáról mind a négy projekthez (Avro, Hive, Impala, Pig) mBox formátumban letöltöttem a havi leveleket. Ezután a Python MboxExtractor moduljával a saját formátumomban tároltam el. A fontos itt is a levelek negyedévenkénti mennyisége volt, így ezekre kellett csoportosítani és összegezni a tárolás során.

Az eltárolt adatokat grafikonokon vizualizáltam. A félévben vizsgált komponensek közül a Hive-nak az eredménye látható alább:.

| Jira tikettek mennyisége havi szinten | Archív levelek mennyisége havi szinten |
|---|---|
|  |  |





A félév végén meghatároztuk a következő félévek célját, ami egy keretrendszer kialakítása, mely képes elemezni a komponensek népszerűséget befolyásoló tényezőket és a jelenlegi egyetlen kimeneti lehetőséget kibővíti különböző variációkkal, például a komponensek népszerűségét táblázatos formában is megjeleníteni és tervben van, hogy egy másik projekt felhasználja az eredményeket.

Ruzsbánszki Nándor
2018.02.06.