**Опис на проектот**

Целта на овој проект е да се развие апликација за обработка на податоци за анализа на историските податоци за акции од македонската берза. Апликацијата се фокусира на стилот и архитектурата на pipeline и филтри за да обезбеди модуларна и скалабилна обработка на податоци. Системот го автоматизира процесот на преземање, трансформирање и чистење на дневните берзиски податоци за сите компании на македонската берза во изминатите 10 години. Ова вклучува преземање податоци од надворешни извори, филтрирање само неопходни информации (како датум, цени на отворање и затворање, обем на тргување и високи/ниски вредности) и нивно форматирање за складирање во база на податоци. Излезот ќе биде база на податоци подготвена за понатамошна анализа во идните задачи, каде што ќе се имплементираат дополнителни стилови на архитектура за да се изгради апликација за анализа на берзата со целосни карактеристики.

Првичната апликација ќе ја автоматизира обработката на податоците, обезбедувајќи висок квалитет и конзистентност на податоците. Оваа задача ја поставува основата за напредна анализа на акции со создавање на сигурен, рационализиран проток на податоци за ефикасно чистење и структурирање на информациите за берзата.

Апликацијата е дизајнирана со серија филтри со помош на архитектурата Pipe and Filter кои се применуваат во низа за прецизно и ефикасно обработување на податоците. Овој пристап осигурува дека секој филтер извршува фокусирана задача, при што излезот од еден филтер беспрекорно поминува во следниот филтер како влез. Во секоја фаза, филтрите ќе бидат одговорни за справување со специфични трансформации, од почетно пребарување на податоци до финално форматирање на податоците. Овој pipeline за податоци, исто така, ќе служи како основа за пообемна примена во идните фази, каде што други софтверски архитектури ќе бидат поставени на врвот за дополнителни функционалности.

**Функциски и нефункциски барања**

**Функциски барања**

* Извлекување податоци од веб-страница за акции од одредена URL адреса користејќи Requests и BeautifulSoup.
* Трансформација на податоци-Апликацијата мора да ги филтрира необработените податоци за да ги задржи само релевантните информации (на пр. датум, цени на отворање/затворање, обем на тргување, високи/ниски вредности). Трансформирајте и форматирајте ги податоците во стандардизирана структура компатибилна со понатамошна анализа и складирање во база на податоци.
* Автоматско пребарување на податоци - Системот мора автоматски да ги враќа дневните податоци за акциите за секој издавач котиран на Македонската берза. Податоците треба да ги опфаќаат последните 10 години и да се ажурираат секојдневно.
* Преземање податоци од структурата HTML (табела) и складирање во структуриран формат користејќи DataFrame.
* Трансформации на датуми и форматирање- функции за форматирање на датуми помеѓу објектите за датум време и претставување на стринговите во формат dd.mm.yyyy.
* Филтрирање податоци - спроведување класи за филтрирање (Filter, CodeDownloaderFilter, SaveDataFilter и DataCompletenessFilter) за обработка и филтрирање на податоците.
* CodeDownloaderFilter - со помош на оваа класа се врши извлекување на податоците од паѓачката листа од веб страницата.
* SaveDataFilter: со оваа класа се проверува дали постојат податоци за секоја компанија и враќа најнови податоци или ги превзема податоците од последните 10 години ако тоа веќе не е направено.
* DataCompletenessFilter: се потврдува комплетноста на податоците, осигурувајќи се дека податоците за секоја компанија се ажурирани до тековниот датум.
* Трансформација на податоци- чистење, обработка и трансформација на преземените податоци, применувајќи конверзии како процентуална нормализација и трансформации на цели броеви за одредени колони.
* Ракување со податоци во Excel- проверка на веќе постоечките датотеки во Excel за секоја компанија, зачувување податоци како датотеки во Excel, додавање нови податоци ако датотеката постои или креирајте нова датотека ако не постои , обезбедувајќи точна агрегација на податоци.

**Нефункциски барања**

* Изведба- операциите за преземање и зачувување податоци треба да се ефикасни, особено при обработка на големи количини на податоци.
* Минимизирање на времето на web-scraping со намалување на интервалите на спиење.
* Приспособливост- дизајнирајте ги филтрите и функциите за обработка за лесно додавање повеќе компании или прилагодување на промените во структурата на HTML на изворната веб-страница.
* Сигурност – справете се со грешките внимателно и пријавувајте ги на соодветен начин, особено при преземање податоци од веб-страници. Апликацијата мора да обезбеди висок квалитет на податоците со проверка на конзистентноста на податоците во различни фази на обработка.
* Обезбедете точност на податоците преку правилни трансформации на датуми, проверки за валидација и постојано складирање на податоци.
* Олеснете го ажурирањето на поединечните филтри без да влијае на целиот тек на обработка на податоци.
* Перформанси - Процесот на пронаоѓање и трансформација на податоците треба да заврши во разумна временска рамка, идеално во рок од неколку секунди до една минута за дневните податоци на секој издавач.
* Осигурете се дека апликацијата е отпорна на погрешно или неочекувано внесување податоци.
* Приспособливост- треба да се справи со зголемените количини на податоци со текот на времето без значително влошување на перформансите.
* Конзистентност на податоците-обезбедете интегритет на податоците помеѓу сесиите за web-scraping, избегнувајќи дупликат записи во датотеките на Excel.Осигурете се дека податоците на секоја компанија се ажурираат доследно, со валидација заснована на датуми за да се спречат преклопувања.

**A black and white ovals with text

Description automatically generated**

**A screenshot of a computer screen

Description automatically generated**

**A close-up of a black background

Description automatically generated**