МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математического и компьютерного моделирования

Разработка системы обработки и хранения данных на основе технологии Big Data

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 247 группы

направление 09.04.03 – Прикладная информатика

механико-математического факультета

Сидоренко Дмитрия Александровича

Научный руководитель
проф., д.ф.-м.н., доцент Д.В. Кондратов

Зав. кафедрой
зав.каф., д. ф. – м. н., доцент Ю.А. Блинков

Саратов 2019

Введение. Колоссальные объемы данных требуют соответствующих технологий. Сегодня компании должны обрабатывать колоссальное количество данных в объемах, которые трудно представить. Это приводит к тому, что традиционные базы данных не могут справиться с такой задачей, и это приводит к необходимости внедрять технологию «Больших данных» (Big Data).

Актуальность данной работы заключается в том, что с каждым годом объём неструктурированных данных увеличивается. Скорость увеличения объёма данных слишком велика и там, где на сегодняшний день объём данных не велик, через пару лет он разрастется до такой степени, что можно будет говорить про Big Data. Благодаря чему актуальность использования технологии Big Data будет расти с той же скоростью, с которой растет и объем данных.

Целью данной работы является разработка научно-технической базы для работы с данными финансового рынка на основе технологии Big Data, что представляет собой хранилище данных, а также методы для обработки и анализа данных.

Для наполнения научно-технической базы, а именно для заполнения базы данных исходными данными и дальнейшей обработки этих данных, с сайта центрального банка Российской Федерации скачена динамика официальных курсов имеющихся валют с 01.01.2000 по 01.01.2018 года.

Для решения поставленной цели необходимо реализовать следующие задачи:

- изучить предметную область, связанную с финансовым рынком;
- изучить технологию Big Data и соответствующие ей методики анализа данных;
- выбрать данные для анализа и наполнения хранилища данных;
- реализовать хранилище данных основываясь на технологии Big Data;
- изучить методы анализа исходных данных;
- произвести анализ данных, имеющихся в хранилище.

Данная магистерская работа состоит из 5 глав:

1. Финансовый рынок;
2. Технология Big Data;
3. Хранилище данных;
4. Анализ временных рядов;
5. Реализация анализа курса валют.

Основное содержание работы. На сегодняшний день внедрение технологии Big Data оказывают влияние на информационные технологии во многих сферах человеческой деятельности.

Если взять в рассмотрение банковскую сферу, то каждый день банкиры сталкиваются с колоссальным объемом информации, которая поступает из бесчисленных источников. Грамотная обработка имеющихся инфопотоков позволит решить практически все ключевые задачи банков: привлечение клиентов, повышение качества услуг, оценка заемщиков, противодействие мошенничеству и др. Повышая скорость и качество формирования отчетности, увеличивая глубину анализа данных, участвуя в противодействии отмыванию незаконных средств, эти технологии помогают банкам соответствовать требованиям регуляторов.

Основные задачи, для которых банки используют технологии анализа «Больших Данных», – это оперативное получение отчетности, скоринг, недопущение проведения сомнительных операций, мошенничества и отмывания денег, а также персонализация предлагаемых клиентам банковских продуктов.

В сущности, понятие «Больших Данных» подразумевает работу с информацией огромного объема и разнообразного состава, постоянно обновляемой и находящейся в разных источниках в целях увеличения эффективности работы, создания новых продуктов и повышения конкурентоспособности.

Под термином «Большие Данные» будем понимать различные инструменты, подходы и методы обработки как структурированных, так и неструктурированных данных для того, чтобы их использовать для конкретных задач и целей, которые должны удовлетворять трём основным признакам (3V):

1. Volume — данные измеряются по величине физического объема документов.

2. Velocity — данные регулярно обновляются, что требует их постоянной обработки.

3. Variety — разнообразные данные могут иметь неоднородные форматы, быть неструктурированными или структурированными частично.

В дальнейшем появились дополнительные признаки «Больших Данных»:

- 4V - добавляется veracity;
- 5V - добавляется viability;
- 6V - добавляется value;
- 7V - добавляется variability.

Во всех признаках подчёркивается основная идея «Больших Данных», что определяющей характеристикой является не только физический объём, но и другие категории, существенные для представления о сложности задачи обработки и анализа данных.

Горизонтальная масштабируемость, которая обеспечивает обработку данных - базовый принцип обработки больших данных. Данные распределены на вычислительные узлы, а обработка происходит без деградации производительности.

Для решения задачи по обработке и хранению данных на основе технологии Big Data реализовано хранилище данных. Хранилище содержит в себе данные финансового рынка, а именно данные по различным курсам валют, имеющихся на сайте центрального банка Российской Федерации. Для наполнения хранилища реализован скрип, который перебирает имеющиеся валюты на сайте центрального банка скачивая динамику официального курса выбранной валюты с заданным временным диапазонам и сохраняет скаченные данные в хранилище без предварительной обработки. Для последующей обработки скаченных данных был выбран временной диапазон с 01.01.2000 по 01.09.2018 года.

Для реализации хранения данных, была спроектирована и реализована структура колоночной NoSQL базы данных HBase для хранения как исходных, необработанных данных, так и для хранения данных, прошедших обработку или результатов проведённых анализов над имеющимися данными.

Учитывая разнообразие исходных данных, для каждого случая будет реализована своя таблица. Например, для курсов валют и драгоценных металлов будут реализованы две таблицы.

Взаимодействие с базой данных и работа с данными происходит с помощью языка программирования Python и соответствующих библиотек для работы с базой данных HBase и анализа данных.

Для работы с данными, находящимися в базе данных, применён паттерн проектирования «Model-View-Controller» (MVC). Основная цель применения MVC состоит в разделении данных и бизнес-логики от визуализации. За счет такого разделения повышается возможность повторного использования программного кода и упрощается поддержка, то есть изменения внешнего вида не отражаются на бизнес-логике.

На основе данного паттерна проектирования реализован класс модели и базового контроллера. В нашем случае, реализация представления является необязательной, так как на данном этапе работа с данными будет проводиться с использованием интерфейса командной строки. Реализованная модель и базовый контроллер являются основными инструментами для работы с данными и основой для реализации графического интерфейса пользователя в независимости от вида реализации (веб или десктоп).

Структура рассматриваемых данных представляет собой исторически накопленные данные, которые представляются временными рядами.

Под временным рядом будем понимать собранный в разные моменты времени статистический материал о значении каких-либо параметров исследуемого процесса. Каждая единица статистического материала называется измерением или отсчётом. Также допустимо называть его уровнем на указанный с ним момент времени.

Уровни ряда – это показатели, числовые значения которых составляют динамический ряд, т.е. они отображают количественную оценку (меру) развития во времени изучаемого явления. Время – это моменты или периоды, к которым относятся уровни.

Во временном ряде для каждого отсчёта должно быть указано время измерения или номер измерения по порядку. Временной ряд существенно отличается от простой выборки данных за счёт того, что при анализе учитывается взаимосвязь измерений со временем, а не только статистическое разнообразие и статистические характеристики выборки.

В анализе временных рядов выделяются две основные задачи:

- задача идентификации;
- задача прогнозирования.

Задача идентификации при анализе временных рядов отвечает за нахождение параметров системы, породившей данный временной ряд — размерности вложения, корреляционной размерности, энтропии.

Цель задачи прогнозирования - по данным наблюдений предсказать будущие значения измеряемых характеристик изучаемого объекта, т.е. составить прогноз на некоторый отрезок времени вперед. На сегодняшний день разработаны и обоснованы различные методы прогноза. Однако все они подразделяются на два основных класса: локальные и глобальные.

Такое деление проводится по области определения параметров аппроксимирующей функции, рекуррентно устанавливающей следующее значение временного ряда по нескольким предыдущим.

На основе имеющихся временных рядов разработан класс, который реализует базовые методы для предварительного анализа и прогнозирования, а именно:

- построение скользящей средней;
- построение экспоненциального сглаживания;
- построение двойного экспоненциального сглаживания - метод Холта;
- построение тройного экспоненциального сглаживания - метод Холта-Винтерса;
- построение кросс-валидации;

- построение модели ARIMA.

Данный класс реализован на высокоуровневым языке программирования общего назначения – Python. Данный язык ориентирован на повышение производительности разработки и читаемости кода. Язык отличается простым синтаксисом, гибкостью в работе и высокой скоростью реализации проектов. Он является одним из основных языков на котором реализуют нейтронные сети, программное обеспечение для анализа Big Data и разработок в сфере искусственного интеллекта, а также из-за своей универсальности и относительной простоты является популярным в научной среде.

За счёт того, что обработка большого объёма информации требует больших затрат ресурсов компьютера. На пример, для реализации построения модели ARIMA весь объём обрабатываемой информации для ускорения процесса обработки и построения прогноза загружается в оперативную память компьютера. Поэтому, для построения данной модели ограничимся данными за последние 3 года.

Анализируемые временные ряды являются не стационарными. Это подтверждается тем, что критерий Дики-Фуллера не отверг нулевую гипотезу о наличии единичного корня. Для приведения исследуемых рядов к стационарному виду воспользуемся преобразованием Бокса-Кокса.

После применения преобразования Бокса-Кокса проверяется критерий Дики-Фуллера. Если критерий не отвергает гипотезу о не стационарности ряда, тогда берутся сезонные разности.

Для определения стационарности ряда, после проведённых преобразований проверяется не только критерий Дики-Фуллера, но и проводится проверка автокорреляционной функции на количество значимых лагов.

Если критерий Дики-Фуллера отвергает нулевую гипотезу о не стационарности, но автокорреляционная функция всё ещё имеет большое количество значимых лагов, тогда берутся первые разности для приведения ряда к стационарному виду.

Исходные ряды, пройдя вышеописанную обработку, становиться стационарным. Итоговые стационарные ряды используются для построения модели ARIMA.

Для построения модели ARIMA по имеющимся стационарным рядам реализован метод, который по автокорреляционной и частной автокорреляционной функции автоматически подбирает параметры для построения модели.

Итогом построения данной модели является график, на котором отображается исходный временной ряд и построенная модель.

Заключение. В ходе выполнения магистерской работы цель была достигнута и для достижения поставленной цели были выполнены следующие задачи:

- изучена предметную область связанная с финансовым рынком;
- изучена технология Big Data, рассмотрено её использование в финансовой отрасли, изучены применяемые в ней методики анализа данных;
- сформирован набор исследуемых данных для анализа и последующего хранения;
- реализовано хранилище данных основываясь на технологии Big Data;
- изучены методы анализа исходных данных;
- произведён анализ данных, имеющихся в хранилище.

Результат данной работы может быть использован для последующего анализа имеющихся данных с целью прогнозирования, выявления наличия и характера корреляционных связей или наличие других особенностей или закономерностей.