Применение ТЕХНОЛОГИИ HADOOP для анализа "больших данных" на примере задачи геоаналитики [[1]](#footnote-1)©

А.Г. Чурсин

Магистрант

Н.А. Тюкачев

Доцент

Д.И. Соломатин

Старший преподаватель

Введение

Данная работа посвящена анализу больших массивов данных с использованием парадигмы MapReduce и стека технологий Hadoop на примере задачи для крупной телекоммуникационной компании, поставщика услуг сотовой связи и широкополосного интернета, в рамках проекта Hadoop for Big Data.

Оператор связи может знать местоположение абонентов. Например, известно, что в такой-то зоне часто появляются люди с iPhone 6, возможно они пользуются стандартом связи 4G, а значит, нужно построить в этой соте LTE-базовую станцию; в другом месте хватит 2G.

Строго говоря, Big Data – это обработка огромных массивов данных, но, например, работа с базой данных под управлением СУБД Oracle на 20 Гб или 4 Гб – это ещё не Big Data, это просто высоконагруженная база данных или highload-БД. Ключевое отличие Big Data от обычных высоконагруженных систем состоит в возможности строить гибкие запросы при обработке данных любого объема.

Одним из самых известных и обсуждаемых проектов в области обработки больших данных является Hadoop, разрабатываемый фондом Apache Software Foundation. Он представляет собой свободно распространяемый набор из утилит и библиотек для разработки и выполнения программ распределенных вычислений [1].

MapReduce – это модель распределенной обработки данных, предложенная компанией Google для обработки больших объёмов данных на компьютерных кластерах [1].

1.  Формулировка постановки задачи

Необходимо сформировать единовременную выгрузку, определив количество уникальных абонентов, которые находились в течение рабочего дня в каждой ячейке зоны разбиения города Нижний Новгород за период с 14 по 20 сентября 2015 года для оптимального расположения офисов монобренда.

Отчет формируется относительно всех зон разбиения (zone id) в дальнейшем именуемых для краткости zid.

Каждые сутки делятся на временные интервалы:

1. 00.00.00-05.59.59;
2. 06.00.00-08.59.59;
3. 09.00.00-20.59.59;
4. 21.00.00-23.59.59.

В поле total (суммарное количество) записывается число абонентов, удовлетворяющих следующему условию. Существует принадлежащая рассматриваемой зоне локация (набор сот), в которой абонент находился некоторое отличное от 0 секунд время в течение рассматриваемого интервала. Если в течение рассматриваемого интервала найдутся несколько зон, для которых это условие выполнено, то абонент приписывается к каждой из них. Таким образом, отчет показывает количество уникальных абонентов, прошедшее через каждую зону в течение рассматриваемого периода.

2.  Анализ требований и обзор входных данных

Требованиями к решению поставленной задачи являются:

1. Предварительный расчет справочника по сетке административного деления г. Нижний Новгород по квадратам 500 на 500 метров за сентябрь 2015 года.
2. Измерение zid должно соответствовать справочнику по сетке деления г. Нижний Новгород на квадраты 500 на 500 метров и содержать все зоны разбиения в результирующем отчете.

Создание конечного программного продукта можно разделить на несколько взаимосвязанных этапов по разработке модулей формирования отдельных вспомогательных продуктов:

1. справочник базовых станций;
2. cправочник покрытия базовых станций и их соседей;
3. cетки административного деления 500 на 500.

Входными источниками являются:

1. справочник с координатами полигонов для Нижнего Новгорода;
2. версионный справочник по секторам свечения базовых станций по регионам;
3. версионный справочник позиций базовых станций по регионам;
4. cправочник с сетками города. Размер ячейки в сетке – 500 на 500.

Наиболее важными при решении поставленных задач являются такие источники как:

1. Трафик Anritsu. Содержит информацию о разобранных событиях сигнальной сети, включая следующие данные: звонки, смс, геолокацию БС. Представляет собой сжатые файлы в формате Sequence File для хранения в HDFS [1]. Источник содержит информацию по транзакциям в сетях 2G и 3G. Периодичность и время появления данных – каждые 15 мин. Средний объем поступающих данных в день – 55 Гб.
2. Трафик Polystar. Схож по содержанию с Anritsu за исключением того, что включает в себя больше событий, больше регионов, а также ошибки сети. Средний объем поступающих данных в день – 12 Гб.

3.  Формирование справочника уникальных базовых станций

Базовая станция (БС) – комплекс радиопередающей аппаратуры (ретрансляторы, приемопередатчики), осуществляющий связь с конечным абонентским устройством – сотовым телефоном.

LAC (Location Area Code) – код зоны местоположения группы базовых станций.

Cell ID – идентификатор конкретной соты внутри Location Area.

БС может быть описана при помощи комбинации LAC и Cell ID.

Для того чтобы рассчитать справочник БС, необходимо объединить файлы справочника по секторам свечения со справочником позиций базовых станций регионов по значению атрибута POSITION\_CODE. В результате получаем файл CITY\_DIM\_BASE\_STATIONS.csv, поля которого, разделенные символом «;», приведены в таблице 1.

Таблица 1

Атрибуты выходных данных справочника БС

| Показатель | Описание измерения |
| --- | --- |
| POSITION\_LATITUDE | Географическая широта |
| POSITION\_LONGITUDE | Географическая долгота |
| CELL | ID соты (номер вышки) |
| AZIMUTH | Направление свечения вышки |
| SECTOR\_ANGLE | Угол свечения вышки |
| RANGE | Cтандарт связи (900/1800/2100 МГц) |

4.  Алгоритм определения свечения вышки на ячейки сетки

На следующем этапе необходимо определить площадь свечения БС. Наносим координаты каждой БС на карту. Координаты БС (значения полей географических широты и долготы) берутся из справочника базовых станций. От получившейся точки откладываем прямую по направлению свечения БС (значение поля AZIMUTH из справочника CITY\_DIM\_BASE\_STATIONS.csv). Затем от получившейся прямой откладываем угол, равный углу свечения вышки. В получившийся угол вписываем окружность с радиусом равным значению, соответствующему стандарту связи. Если для БС угол свечения строго больше 180 градусов, то около вышки описываем окружность радиусом 1 км, затем данную окружность вписываем в треугольник. Дальность свечения БС равна высоте треугольника. Если в справочнике по секторам свечения БС отсутствуют азимут и угол свечения, то вместо площади свечения будет использоваться круг. В дальнейшем получаем справочник покрытия с названием CITY\_DIM\_TRIANGLES\_WKT.csv.

Формируем справочник по сетке 500 на 500. Используем заготовленную сетку по городу в виде .shp-файла – CITY\_GRID\_500.shp. На сетку из справочника города наносим покрытие БС в виде треугольников, получившееся в результате расчета CITY\_DIM\_TRIANGLES\_WKT.csv. Если площади треугольников вышек пересекаются, то они являются соседними.

Далее определяем, на какие ячейки сетки светит данная БС. На рис. 1 приведен пример расположения БС и ее свечения на соседние квадраты сетки города:



*Рис. 1.* Расположение БС и засветы на квадраты

Рассчитываем площадь свечения вышки в каждой ячейке сетки (2, 3, 4, 6, 7, 8, 10, 11, 12) и заносим в справочник DIM\_GRID\_500.csv. Атрибуты справочника приведены в соответствии с содержимым таблицы 2:

Таблица 2

Атрибуты выходных данных справочника DIM\_GRID\_500.csv

| Показатель | Описание показателя |
| --- | --- |
| ID\_SQUARE | Идентификатор квадрата, на которую светит данная вышка |
| LAC | Номер группы вышек. |
| CELL\_ID | ID соты (номер вышки). |
| AREA\_GLARE | Площадь засвета квадрата, получившегося в результате пересечения квадрата и сектора вышки |
| AREA\_COVERAGE | Площадь фигуры полученной в результате пересечения треугольника (покрытия БС) и квадрата (ячейки сетки) |

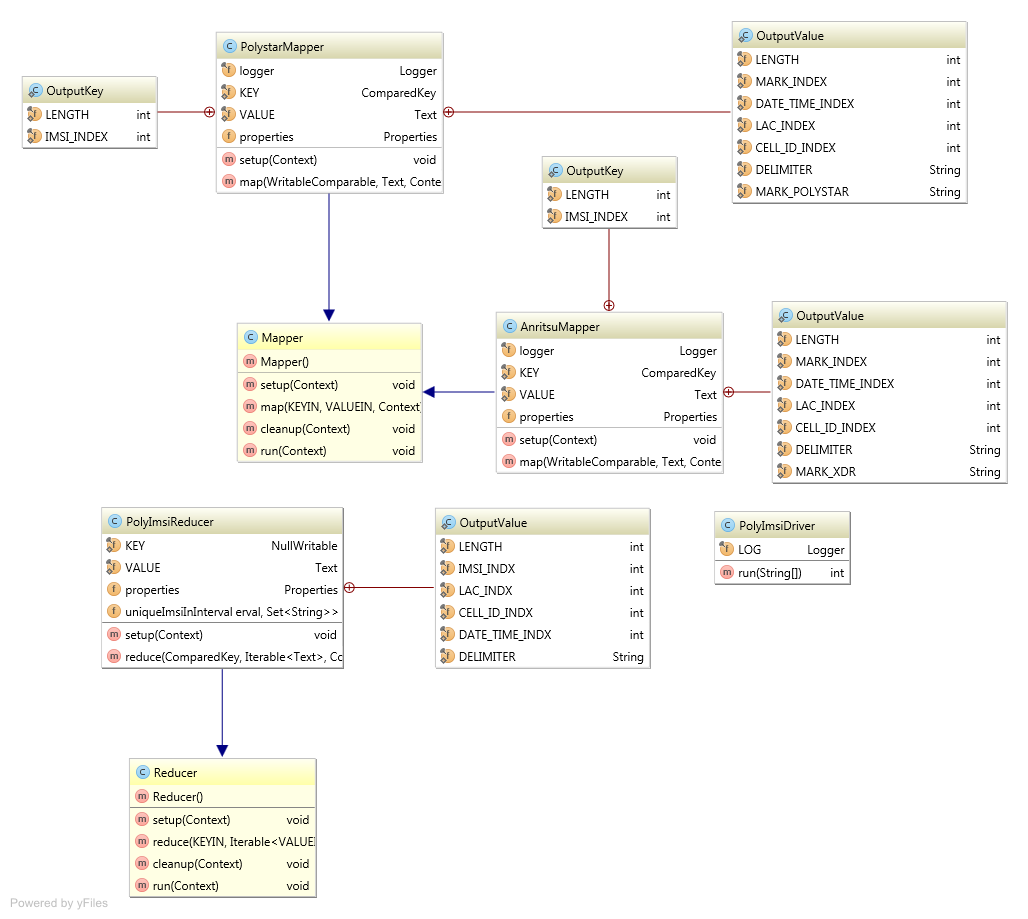
5.  Реализация модуля «Монобренд» и анализ полученного результата

Данный раздел описывает основные особенности непосредственной разработки и анализа полученных данных в результате эксперимента.

Разработку основного модуля «Монобренд» условно можно разделить на 3 этапа (по количеству stage-пакетов):

1. фильтрация данных источников Anritsu и Polystar (пакет polyimsi);
2. фильтрация данных справочника DIM\_GRID\_500.csv (пакет grid) и соединение данных, полученных на выходе первой стадии с отфильтрованными данными;
3. группировка и подсчет значений требуемых показателей (пакет groupby).

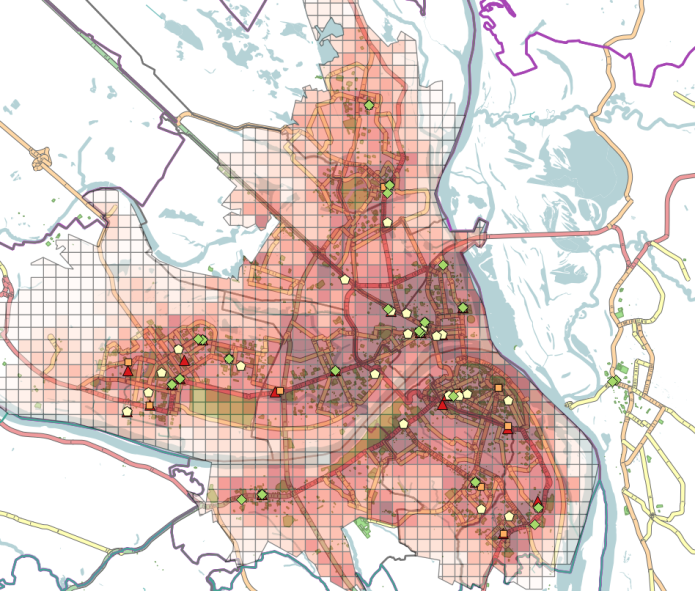
Для первого этапа разработаны основные классы для получения событий сигнальной сети из источников – AnritsuMapper, PolystarMapper, PolyImsiReducer, реализующие абстрактные классы Mapper и Reducer из стандартного набора Hadoop-библиотек второй версии. Диаграмма классов представлена на рис. 2.



*Рис. 2.* Диаграмма классов пакета polyimsi модуля «Монобренд»

Для последующего анализа полученного результата потребовалось программное обеспечение QGIS Desktop 2.8.4 [2]. На основе имеющихся .shp-файлов строится карта города. Затем результирующий отчет привязывается к топологии по имеющимся координатам. Таким образом, мы получаем нечто похожее на тепловую карту с указанием распределения полученных количественных показателей. Параметры цветовой гаммы настраиваются внутри программы QGIS по процентилям [2].

На представленном ниже рис. 3 отражена ситуация за 14 сентября 2015 года, а именно количество активных абонентов в промежуток времени с 9 утра до 21 вечера. Для обозначения точек продаж разных мобильных операторов, включая компании конкурентов, приняты условные обозначения. Они приведены как пояснения к рис. 3.



*Фигуры на карте: «Билайн» – треугольники; «Мегафон» – квадраты; «МТС» – пятиугольники; «Теле2» – ромбы.*

*Рис. 3.* Количество активных абонентов в городе Нижний Новгород в промежуток с 9 до 21 часов после полудня

Заключение

В ходе работы над задачей геоаналитики были разработаны все основные и вспомогательные модули, а именно:

1. модуль формирования справочника базовых станций;
2. модуль формирования cправочника покрытия базовых станций и их соседей;
3. модуль формирования cетки административного деления 500 на 500;
4. модуль формирования результирующего отчета «Монобренд».

Результат работы программ, а именно полученный отчет, соответствуют критериям качества со стороны компании-заказчика, все программные модули прошли тестирование и приняты как продуктивные задачи для автоматизированной работы на промышленном кластере.

Список литературы

1. White, T. Hadoop: The Definitive Guide, 2nd Edition / T. White. – Sebastopol : O’Reilly Media, 2010. – 628 p.
2. Документация QGIS 2.0 [Электронный ресурс] : оф. сайт. – Режим доступа : <http://docs.qgis.org/2.0/ru/docs/index.html>

1. © Чурсин А.Г., Тюкачев Н.А., Соломатин Д.И., 2016 [↑](#footnote-ref-1)