# Общее задание

1. Решить задачу с помощью Java-программы на MapReduce
   1. записать суммарное время обработки данных (CPU time).

*Замерять время работы задачи в общем случае не имеет смысла, так как параллельно на кластере часть ресурсов может быть выделено под другие задачи других пользователей.*

* 1. предложить шаги по оптимизации обработки данных (например, уменьшение количества этапов Map-Reduce, оптимизация Join)
  2. реализовать оптимизацию обработки данных, записав суммарное время обработки.

1. Решить задачу на Hive SQL.
   1. Провалидировать получившийся датасет с результатом MapReduce
   2. Записать суммарное время обработки данных
   3. Сравнить результаты с Java MapReduce, объяснить разницу в скорости работы MapReduce и Hive
   4. Попробовать запустить те же SQL запросы на Impala SQL, описать результат (как положительный, так и отрицательный)
2. Решить задачу на Apache Spark (Java 8 или Scala)
   1. Провалидировать получившийся датасет с результатом MapReduce
   2. Записать суммарное время обработки данных
   3. Сравнить результаты с Java MapReduce и Apache Hive, объяснить разницу в скорости работы
   4. Перечислить возникшие проблемы и субъективно сравнить сложность реализации с предыдущими методами

## Предлагаемые шаги

1. Просмотреть исходные структуры данных, пути где они лежат и выписать используемые файлы и атрибуты. Просмотреть примеры значений исходных атрибутов, для того, чтобы оценить качество данных и необходимость их очистки или преобразования (trim(), toLowerCase()). Документацию по исходной схеме Stackoverflow можно посмотреть тут: <https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede>

При желании, к данным основного сайта stackoverflow.com, можно добавить другие сайты из семейства stackexchange. Все они имеют одинаковые структуры данных, поэтому логика обработки меняться не будет.

1. Составить на бумаге план выполнения задачи в простых шагах (группировка, сортировка, join)
2. Реализовать шаги, проверяя корректность данных после каждого шага, например, с помощью <https://data.stackexchange.com/stackoverflow/query/new> или другой технологии (Hive, Impala). Для больших таблиц можно использовать подмножество данных (например, Posts\_sample) для экономии времени.
3. После того, как реализация будет проверена на подмножестве данных, можно запустить ее на полном наборе данных, замерив производительность.

# Найти Top 10 трендовых технологий

Технологии в IT-области развиваются достаточно быстро, поэтому IT-специалисты должны постоянно следить за трендами и изучать новые технологии, фреймворки, паттерны, подходы. С другой стороны, работодатели, которые заинтересованы в поиске лучших специалистов, будут стремиться использовать наиболее трендовые технологии, чтобы повысить привлекательность вакансии.

Предполагается, что популярность технологии отражается в степени активности пользователей StackOverflow. Вопрос ассоциируется с технологией по наличию тега.

Для определения популярности, можно использовать (с разными весами):

* наличие вопроса с тегом технологии
* количество ответов
* количество голосов за вопрос
* активность в комментариях

Увеличение популярности можно понимать как увеличение доли вопросов/ответов/комментариев со временем.

Например, можно взять одинаковые интервалы (по годам или месяцам), определяя прирост популярности, как процент от кол-ва вопросов предыдущего периода. Таким образом, самые трендовые технологии можно определить, отсортировав по сумме прироста популярности в течение нескольких последних лет.

# Найти технологии, которые не устаревают

Некоторые технологии со временем прочно занимают свою нишу (выходят на «плато продуктивности» <https://www.gartner.com/en/research/methodologies/gartner-hype-cycle>) и могут оставаться актуальными в течение десятков лет.

Предполагается, что популярность технологии отражается в активности пользователей StackOverflow. Вопрос ассоциируется с технологией по наличию тега.

Стабильную популярность технологии можно понимать как постоянное число вопросов/ответов за одинаковые интервалы времени. Возможно незначительное уменьшение вопросов/ответов, если уже была накоплена большая база знаний. Например, если взять одинаковые интервалы (по годам или месяцам), определяя колебания популярности, как процент от кол-ва вопросов предыдущего периода.

Для определения популярности, можно использовать (с разными весами):

* наличие вопроса с тегом технологии
* количество ответов
* количество голосов за вопрос
* активность в комментариях

*Пример расчета:* взять одинаковые интервалы (по годам или месяцам), определяя изменение популярности, как процент от кол-ва вопросов предыдущего периода. Таким образом, неустаревающие технологии можно определить как те, сумма процентов изменения популярности которых по интервалам времени минимальна в течение всей истории stackoverflow.

Если предположить, что модель Gartner-а верна для исследуемых технологий, то график популярности должен повторять форму кривой Hype Cycle (возможно, с несколькими «горбами»). В этом случае, стабильность технологии имеет смысл начинать измерять только с выхода на «плато продуктивности».

# Найти пользователей, которые быстрее всех правильно отвечают на вопросы

Назовем временем реации пользователя разницу (например, в секундах) между заданным вопросом и ответом, используя даты публикации постов типа «вопрос» и «ответ». При этом надо учитывать только правильные ответы. Среднее время реакции считать как среднее по всем постам, сгруппированным по автору ответа.

Требуется вывести список пользователей (id, имя), отсортированных по времени реакции по возрастанию.

В исходных данных много незаполненных полей, например, есть user id, нет user name в Posts. Имя и другие атрибуты пользователей лучше брать из Users.xml

Возможны исключительные случаи, например, время ответа с точностью до секунды совпадает со временем создания вопроса. Возможно, это боты, накручивающие репутацию пользователя (так как на stackoverflow многие работодатели ищут специалистов). Такие случаи имеет смысл отфильтровать из выборки.

# Найти самых лучших специалистов по Hadoop в России

Рекрутеры часто используют специализированные профессиональные сайты для поиска кандидатов с требуемым профилем технологий. Компании нужно найти потенциальных кандидатов в России, которые хорошо знают стек Apache Hadoop.

«Лучшими» специалистами можно считать пользователей с наибольшим рейтингом (Score). В простейшем варианте, можно использовать рейтинг самого пользователя (из Users.xml) и смотреть есть ли у него ответы на вопросы с тегами “Hadoop”. Но туда могут попасть пользователи которые, например, хорошо отвечали на множество вопросов по Java, а на Hadoop ответили только на 1 вопрос. Более честный вариант был бы только тот, который учитывал рейтинг ответов по тегу “Hadoop”.

Поле Location в users – текстовое поле со свободным вводом любого текста. Поэтому, для определения того, находится ли пользователь в России, придется провести анализ данных поля Location и определять это по наличию подстрок “Russia”, “Moscow”, “Russian Federation” и др. Для этого можно вывести список уникальных Location, отсортированный по сумме рейтингов и просмотреть глазами верхние N строк, которые покрывают большую часть кандидатов.

Вместо Hadoop, можно взять любую другую распространенную технологию: Java, Spark, JavaScript, С++, Python

# 5. Найти страны, в которых наиболее активно развивается IT

За метрику “развития IT” можно принять, например, изменение активности (количество заданных вопросов, ответов и комментариев) пользователей определенной страны за последние годы.

Поле Location в users – текстовое поле со свободным вводом любого текста. Поэтому, для определения страны, придется провести анализ данных поля Location и определять это по наличию подстрок с названием страны (список стран взять из Википедии). Для этого можно вывести список уникальных Location, отсортированный по сумме рейтингов и просмотреть глазами верхние N строк, которые покрывают большую часть кандидатов.

Для некоторых стран (в том числе для России), есть отдельные сайты stackoverflow (см. <https://archive.org/download/stackexchange>), которые тоже имеет смысл учитывать в статистике.