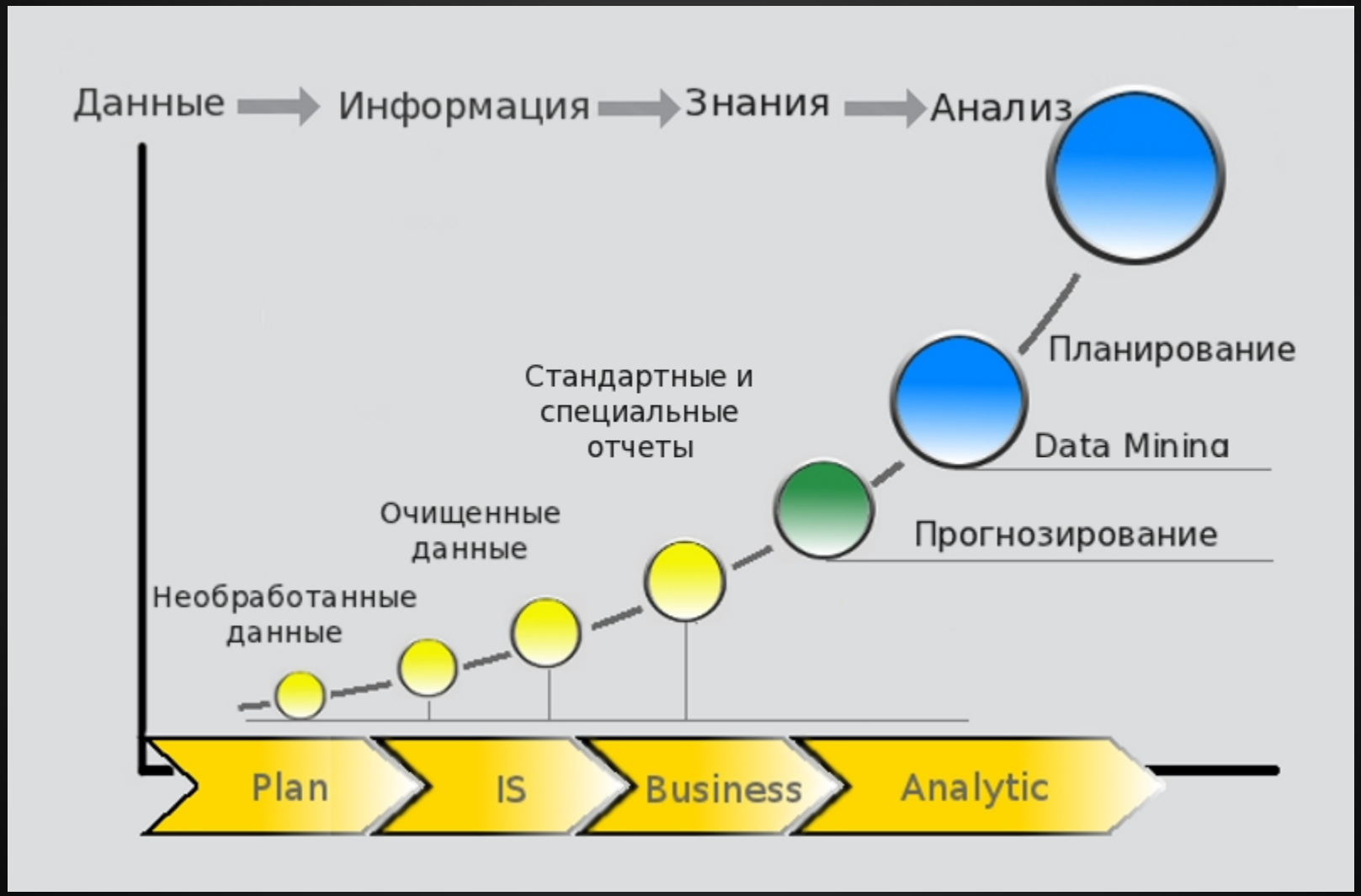
Оптимизация хранения и обработки большого объема статистических данных на уровне хранилища данных

Студент: Марченко С.В.
Руководитель: Астахов А.А.

Кафедра математических и информационных технологий
Санкт-Петербургский Академический Университет
2012

Система поддержки принятия решений



Архитектура аналитической системы

Задачи исследования

- сформулировать требования к системе для работы с СД
- рассмотреть технологии хранилищ данных
- провести сравнительный анализ систем хранения
- интерпретировать результаты анализа
- определить оптимальный способ работы с СД

Структурные элементы исследования

- реализация оптимизированных вариантов хранения и обработки данных
- разработка инструмента для запуска тестов
- запуск тестовых задач на реальных данных
- формирование набора рекомендаций для практического применения

Сравнительный анализ

Основной подход:

NoSQL технологии (Apache Hadoop)

Тестируемые хранилища:

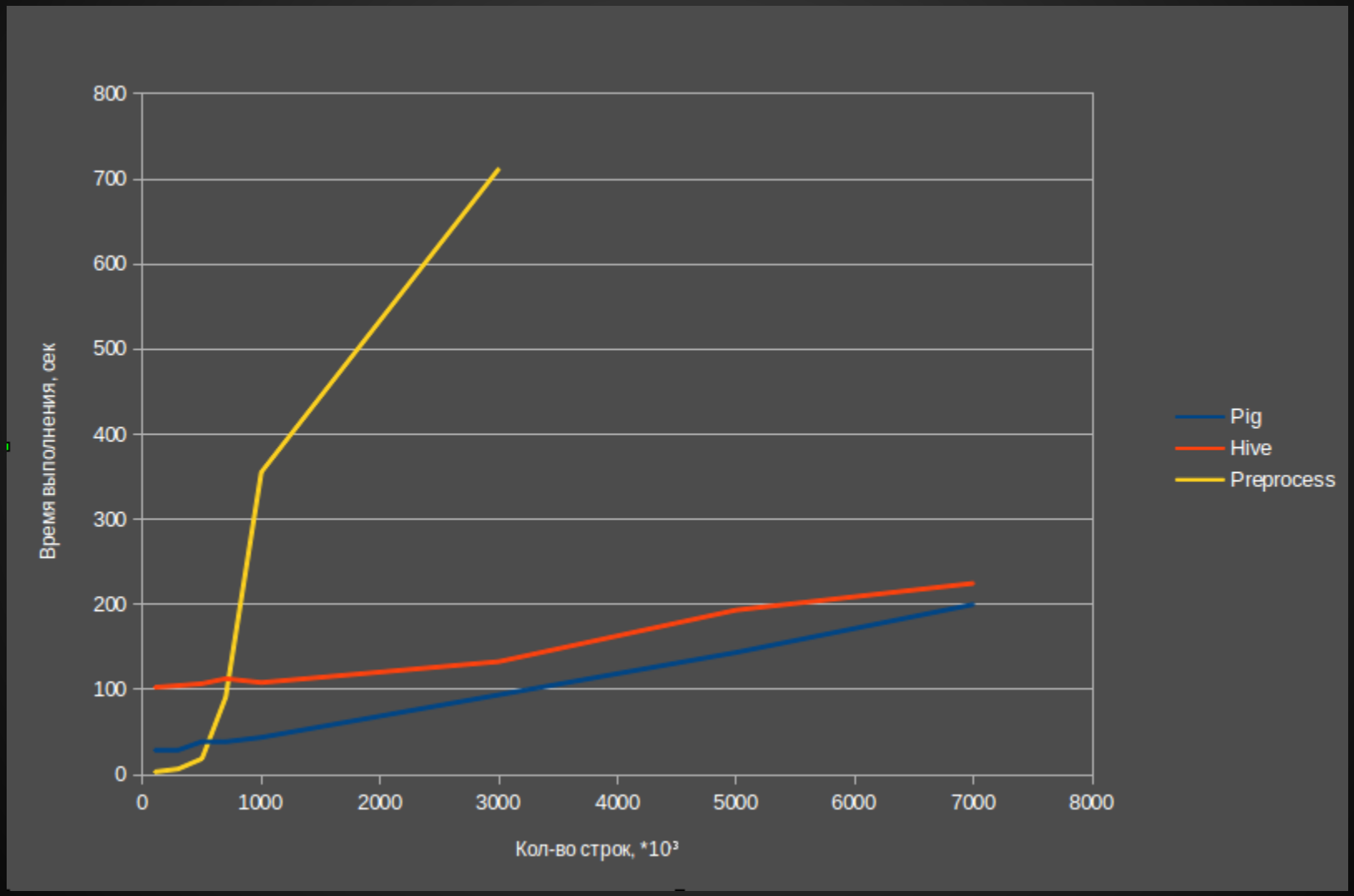
- Hadoop Hive
- Hadoop Pig + HDFS
- Preprocessing + MySQL

Сценарии тестирования

Единая шкала оценки:

- задачи
  - пакетная загрузка данных
  - агрегация
  - join (пересечение данных)
  - доступ к данным
- первичные данные
- аппаратная платформа

Результаты замеров



Инструмент запуска задач

- нет привязки к конкретной технологии хранилища данных
- API для логгирования результатов
- API визуализации замеров
- интерфейсы для загрузки данных, запуска тестов и задач
- план запуска

План запуска

Итоги сравнительного анализа

Набор рекомендаций относительно:

- структуры хранения данных
- плана выполнения запросов
- средств обработки (Hadoop)

Спасибо за внимание!

Вопросы?