Выполнение декларативных сценариев аналитической обработки данных на основе оптимизации и приближенного вычисления

Анна Ярыгина, Борис Новиков

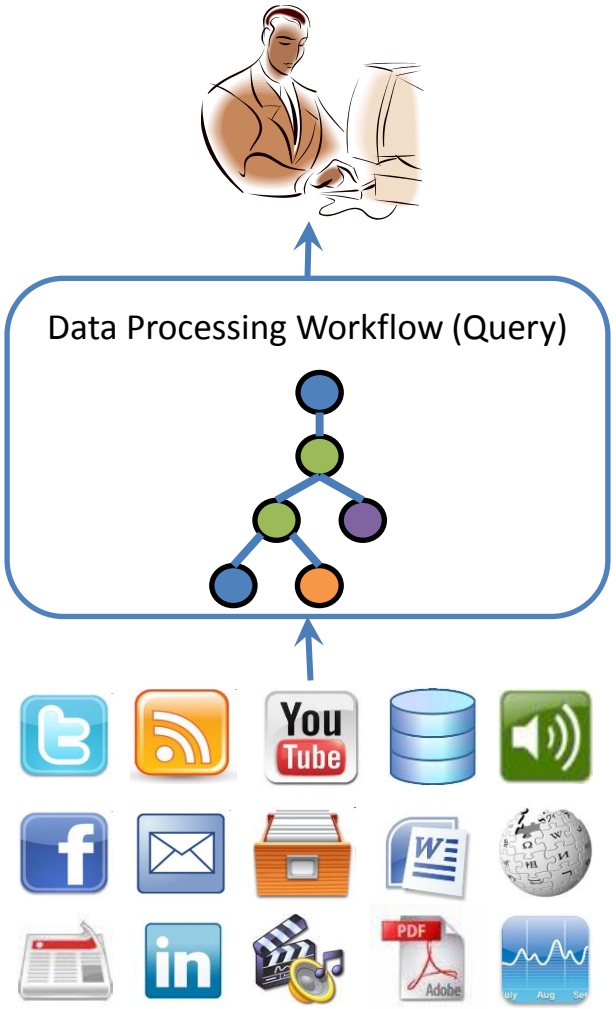
Санкт-Петербургский государственный университет

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Пересмотр подходов

- Объем (Volume)
  - Декларативные языки запросов
  - Распределенное выполнение
- Разнообразие (Variety)
  - Данных
    - Структурированные, неструктурированные
    - Статические, динамические
    - Задача отображения схем
  - Запросов
    - Поиск (данные возвращаются пользователю)
    - Статическая аналитика (данные загружаются в хранилище)
    - Динамическая аналитика (данные передаются в аналитический инструмент)
- Скорость (Velocity)
  - Оптимизация
  - Приближенное выполнение
- Качество (Veracity, Validity)



Системы и языки

Системы

- Параллельные вычисления
  - ASTERIX
  - Scope
  - Hive
- Оптимизация и выполнение сложных аналитических сценариев
- Выполнение специфических классов запросов при ограниченных ресурсах (приближенное выполнение на основе выборок)
  - Система приближенного выполнения SQL запросов с агрегированием
  - Система исследования данных на основе приближенного выполнения SQL запросов

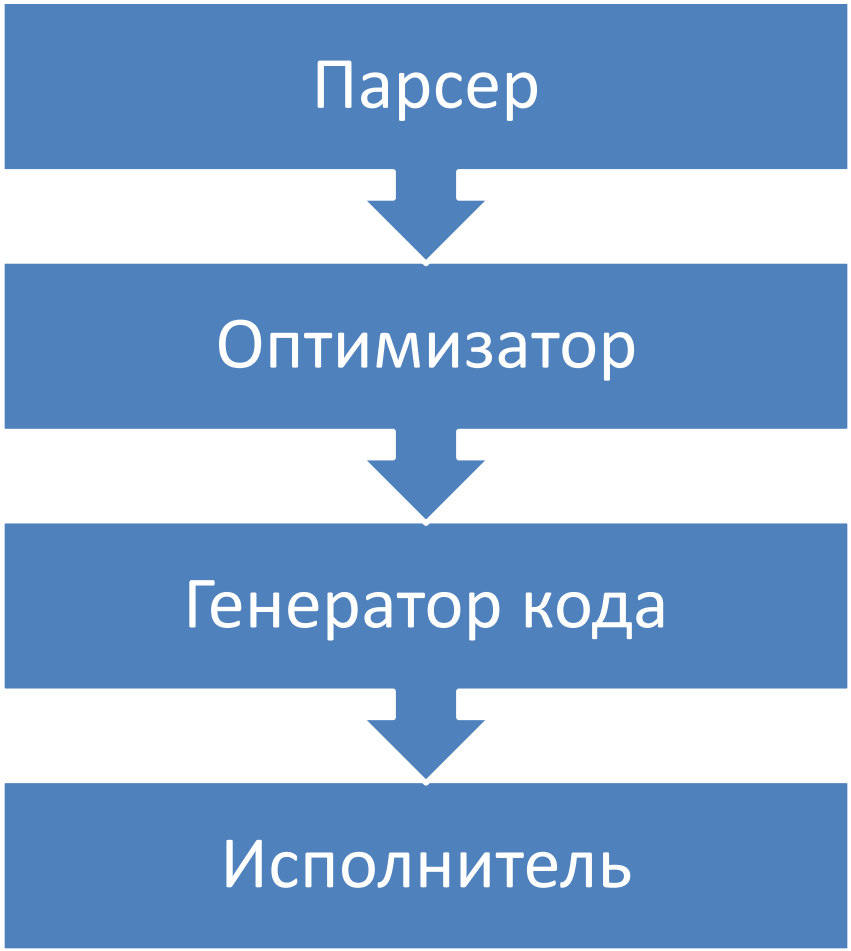
Языки

- Algebricks алгебраический слой для оптимизации и параллельного выполнения запросов
- Язык Hyracks
- Язык Scope
- Язык Pig Latin
- Язык HiveQL

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Выполнение точных запросов



- Декларативные языки запросов
- Реляционная алгебра
- Оптимизация
  - Алгебраические тождества
  - Алгоритмы выполнения операций
- Потоковое выполнение

Оптимизация запросов

Задача оптимизации

- Найти план выполнения запроса минимизирующий функцию стоимости

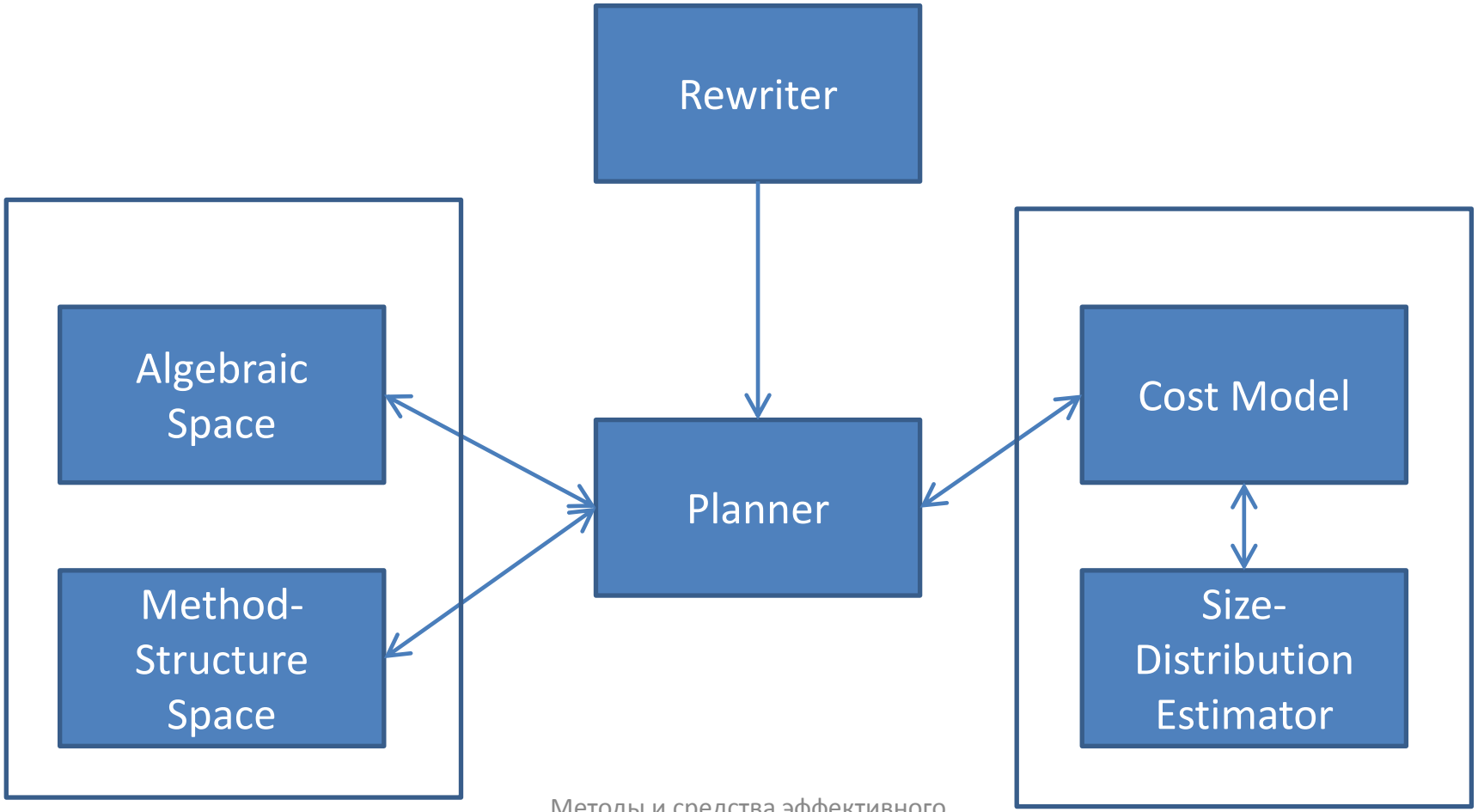
Функция стоимости

- Время выполнения запроса
- Время получения первого кортежа ответа
- Процессорное время
- Объем ввода/вывода

Ресурсы

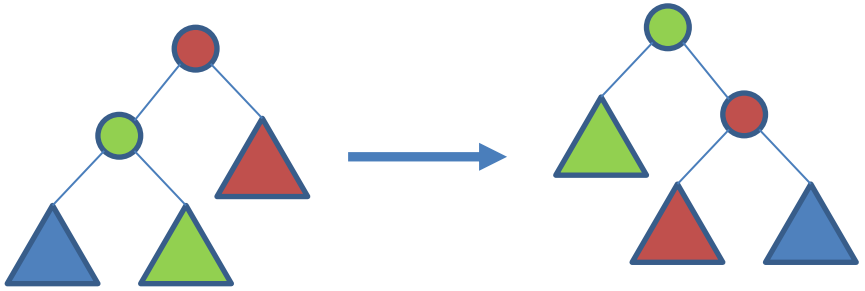
- Стоимость и ресурсы
- Сложность и конфигурация
- Аддитивный ресурс
  - Распределяемый между операциями
  - Стоимость сценария определяется суммой ресурсов, выделенных операциям

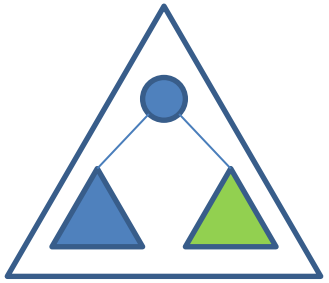
Архитектура оптимизатора



Алгоритмы поиска плана

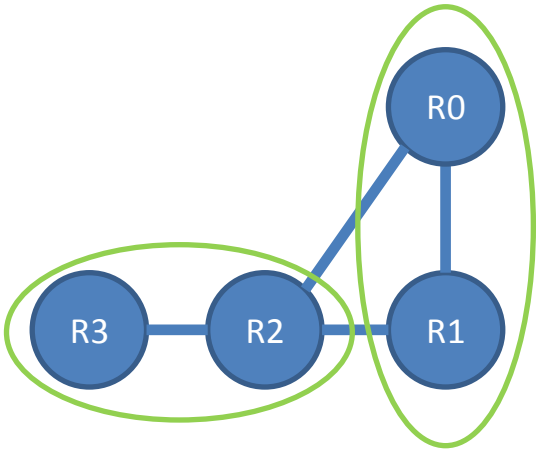
- Сверху вниз
  - Метод случайного поиска
  - Генетические алгоритмы
  - Метод ветвей и границ
- Снизу вверх
  - Алгоритм динамического программирования
  - Жадный алгоритм

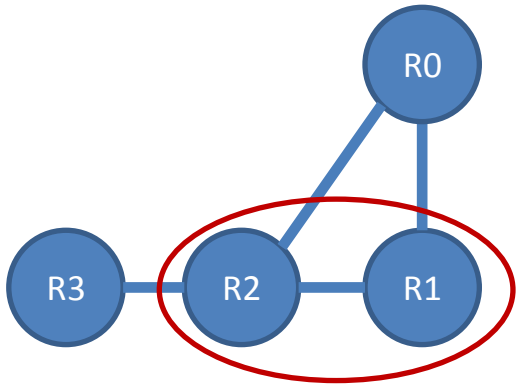




Метод ветвей и границ на графах

- Граф запроса
  - Вершины - отношения
  - Дуги - предикаты
- Идеи
  - Исключение прямые произведений
  - Некоммутативные операции
- Перечисление планов
  - Рекурсивно перечисляем множество всех разбиений
    - Выбираем вершину
    - Рассматриваем ближайших соседей
- Гиперграф запроса





План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Адаптивное выполнение запросов

- Цикл адаптивного выполнения запросов
  - Измерение
  - Анализ и оптимизация
  - Выполнение
- Запросов
  - Потоковое выполнение
  - Промежуточная материализация
- Операций

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Понятие приближенности

Природа приближенности

- История потока
  - Бесконечность истории
- Природа данных
  - Надежность источников
- Природа задачи
  - Поиск по подобию
- Объем данных и скорость обработки

Понятия приближенности

- Приближенное выполнение запросов
- Нечеткие запросы

Расширенные алгебры

Расширение реляционной модели

- Пользовательские веса
- Нечеткие множества
- Фиксированный класс данных

Уровни расширения алгебры

- Работа со списками, оценками, рангами
- Поиск лучших объектов
- Поддержка рангов и оценок внутри операций

Найти 5 лучших отелей согласно рейтингам booking.com и TripAdvisor, если рейтинг последнего в 2 раза важнее.

Расширения операций

- Теоретико-множественные
  - Комбинирование оценок
  - Пользовательские веса
  - *Найти изображения красных и полосатых насекомых*
- Выборка
  - Нечеткий предикат
  - *Найти красных насекомых*
- Проекция
  - Оценки дубликатов
- Соединение
  - Комбинирование оценок
  - Нечеткий предикат
  - *Найти недорогой дом и хорошую школу в одном районе*
- Поиск лучших
  - *Найти 10 самых ядовитых насекомых*
- Выборка по порогу
  - *Найти отели с оценкой качества сервиса не ниже 5*

Особенности оптимизации запросов в расширенных языках

Алгебраические тождества

- Поиск лучших объектов и соединение
- Выборка по значению оценки и соединение
- Симметричное соединение

Модели стоимости

- Оценка селективности
- Оценка размера входных данных
- Тяжелые предикаты

Приближенное выполнение

- Приближенные алгоритмы
- Ограничения на ресурсы
- Any-time

- Приближенные алгоритмы
  - Агрегирование
    - Any-time
    - Оценка качества
  - Соединение на основе подобия
    - Вычисление предиката
    - Доступ к источнику

Качество

- Точность
- Относительный размер выборки

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Многокритериальная оптимизация

- Среди множества планов выполнения запросов выбрать тот, который минимизирует стоимость, вычисленную на основе совокупности критериев
  - время выполнения запроса
  - денежные затраты
  - потеря качества в результате

Параметрическая оптимизация

- Стоимость плана выполнения запроса зависит от параметров с неизвестными в момент оптимизации запроса значениями
- Множество планов, в котором для всех возможных значений параметров запроса имеется хотя бы один оптимальный план

Многокритериальная параметрическая оптимизация

- Понятия
  - Доминирование
  - Регион Парето
  - Множество планов Парето
  - Регион релевантности
- Алгоритм сокращения регионов релевантности
  - Динамическое программирование
  - Сокращение региона релевантности при серии сравнений плана с эквивалентными ему
  - Отбрасывание планов с пустыми регионами релевантности
- Кусочно-линейные функции стоимости (политопы)

Направления развития

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Цель и подходы

Цели

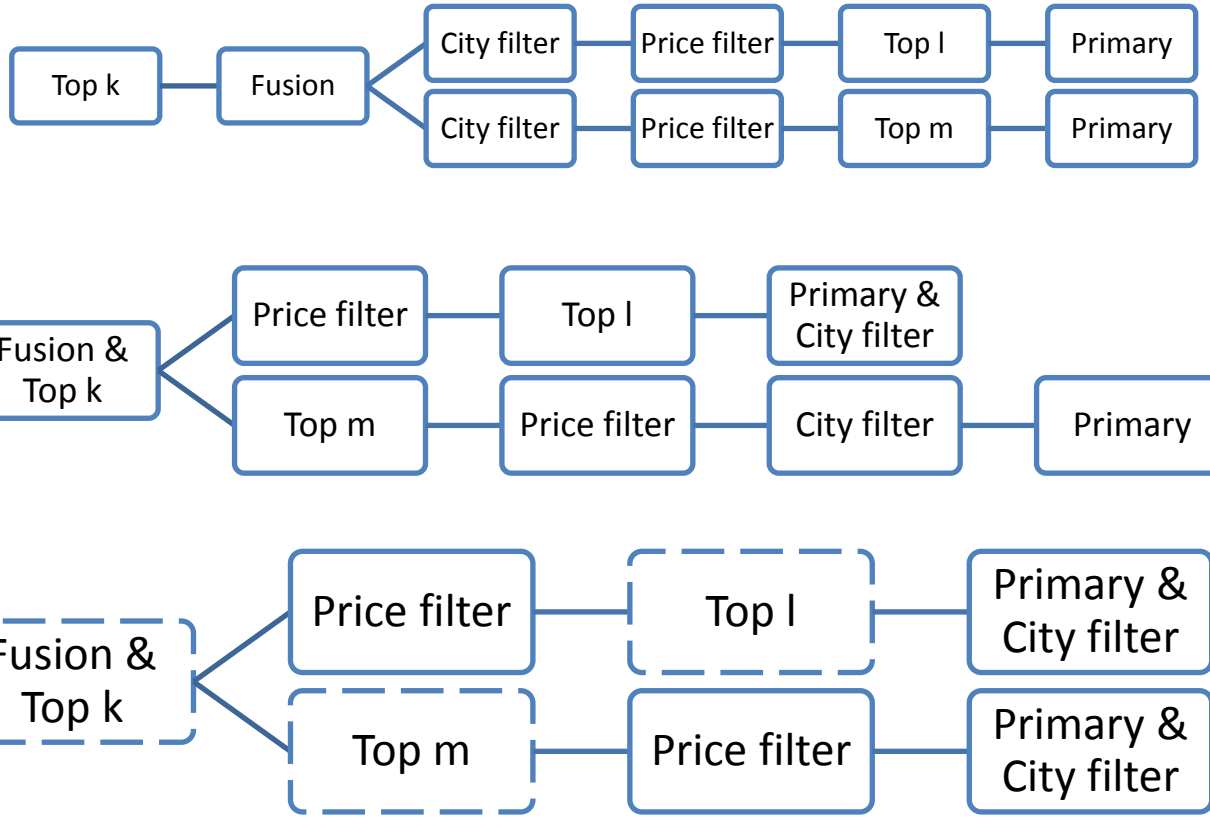
- Единообразно формулировать запросы к данным и описывать их обработку
- Эффективно выполнять сложные запросы
- Выполнение запросов в реальном времени (предсказуемое и контролируемое время ответа)

Предлагаемые подходы

- Набор объектов с оценками
- Расширяемая алгебра операций
- Расширенная модель стоимости (качества)
- Алгоритм распределения ресурсов
- Алгоритмы многокритериальной оптимизации

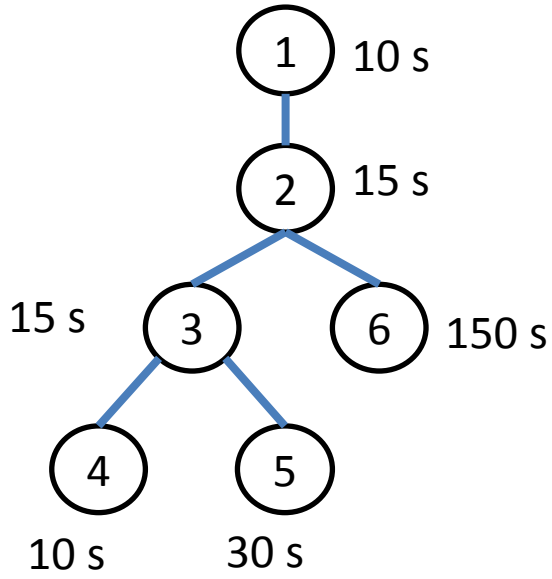
Как оптимизировать?

Find best 10 hotels in London according to service and cleanliness



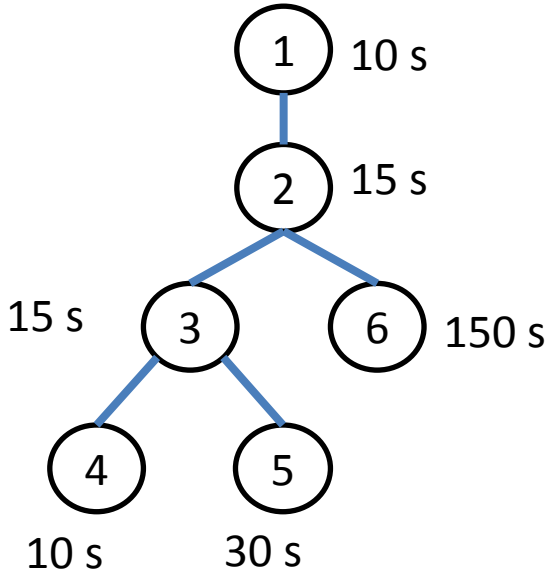
Пример: анализ в реальном времени

1 group by PriceRange nest (ProductGroup, avg(Rating), avg(Sentiment))
2 group join on ProductModel
3 group join on ProductModel
4 get database product table
5 get retailer ratings for products (500 per second)
6 get sentiments from tweets for products (300 per second)

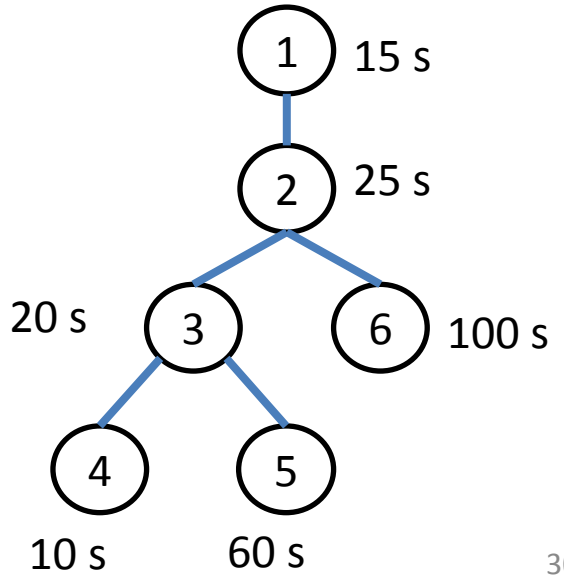


		Product Groups									
		Cam		Scan		Print		Fax		MF	
Price range	< 10%	3,14	3,63	4,06	1,67	1,87	4,71	0,17	1,59	1,71	
	10 – 20%	0,48	2,82	4,85	2,66	4,51	3,40	4,23		4,77	2,08
	20-35%		1,36	2,29	4,43	1,83	3,34	3,04	1,82	4,08	2,22
	35-60%	4,70	0,97	1,56	0,24		3,91	2,65	0,68	2,43	2,30
	> 60%	2,99	1,98	0,31	4,56	0,45	0,63	1,35	1,76	0,32	2,10

Влияние распределения ресурсов



		Product Groups									
		Cam		Scan		Print		Fax		MF	
Price range	< 10%	3,14	3,63	4,06	1,67	1,87	4,71	0,17	1,59	1,71	
	10 – 20%	0,48	2,82	4,85	2,66	4,51	3,40	4,23		4,77	2,08
	20-35%		1,36	2,29	4,43	1,83	3,34	3,04	1,82	4,08	2,22
	35-60%	4,70	0,97	1,56	0,24		3,91	2,65	0,68	2,43	2,30
	> 60%	2,99	1,98	0,31	4,56	0,45	0,63	1,35	1,76	0,32	2,10



		Product Groups									
		Cam		Scan		Print		Fax		MF	
Price range	< 10%	3,21	3,73	4,09	1,71	1,94	4,74	0,17	1,61	1,80	4,39
	10 – 20%	0,54	2,88	4,88	2,72	4,60	3,48	4,28		4,79	2,12
	20-35%		1,41	2,37	4,45	1,90	3,36	3,10	1,89	4,08	2,25
	35-60%	4,74	1,03	1,64	0,28	3,42	4,00	2,70	0,75	2,52	2,40
	> 60%	3,03	2,07	0,41	4,59	0,53	0,71	1,41	1,81	0,33	2,16

Возможные применения

- ETL процессы
- OLAP запросы
- Непрерывный анализ потоков
- Анализ в реальном времени

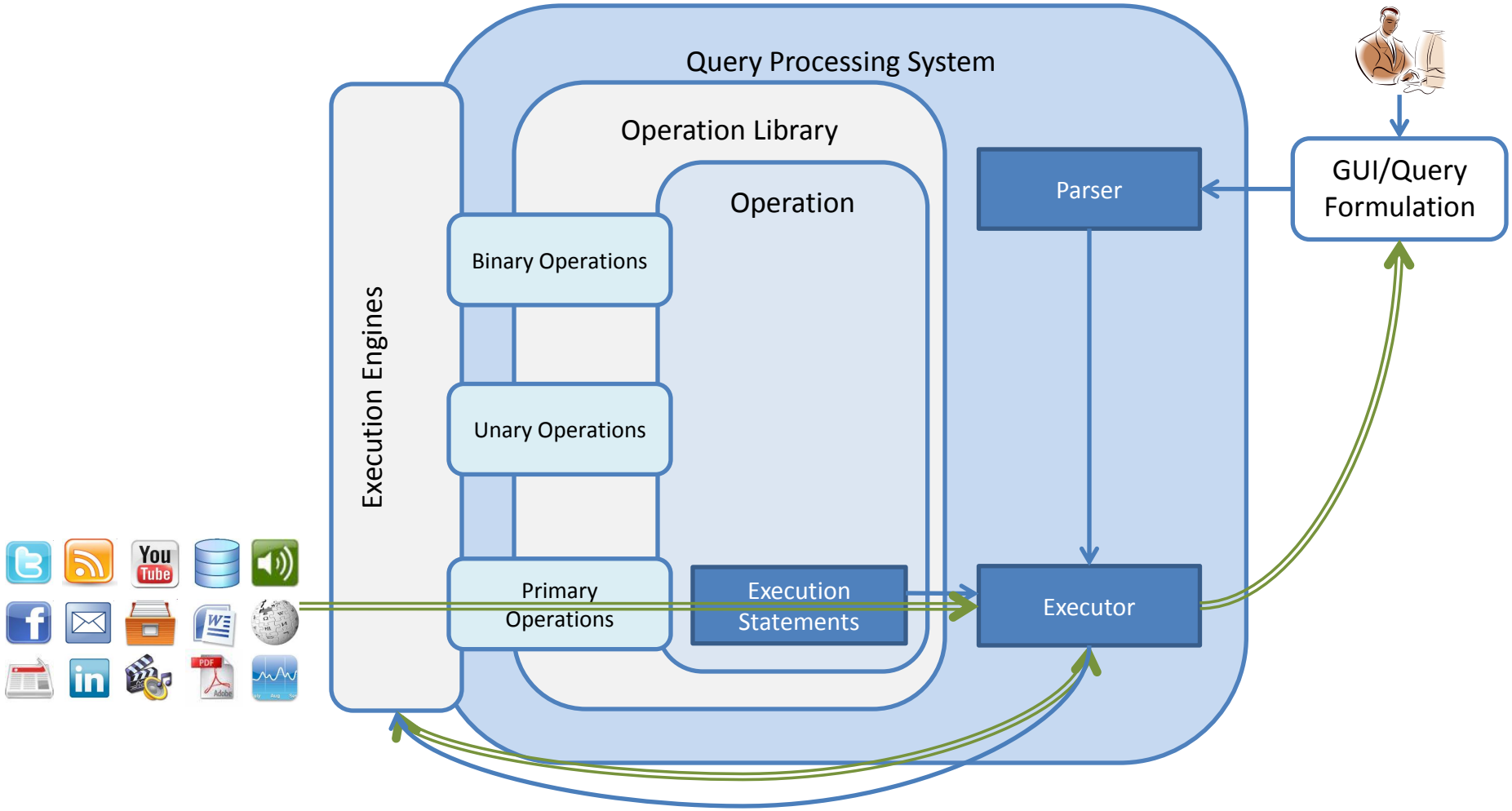
План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

Обзор системы

- Язык и модель данных
- Оптимизация
- Распределение ресурсов
- Исполнение

Предварительная архитектура



Открытая алгебра

Q-set (q,B,S) множество объектов с оценками над базовым множеством

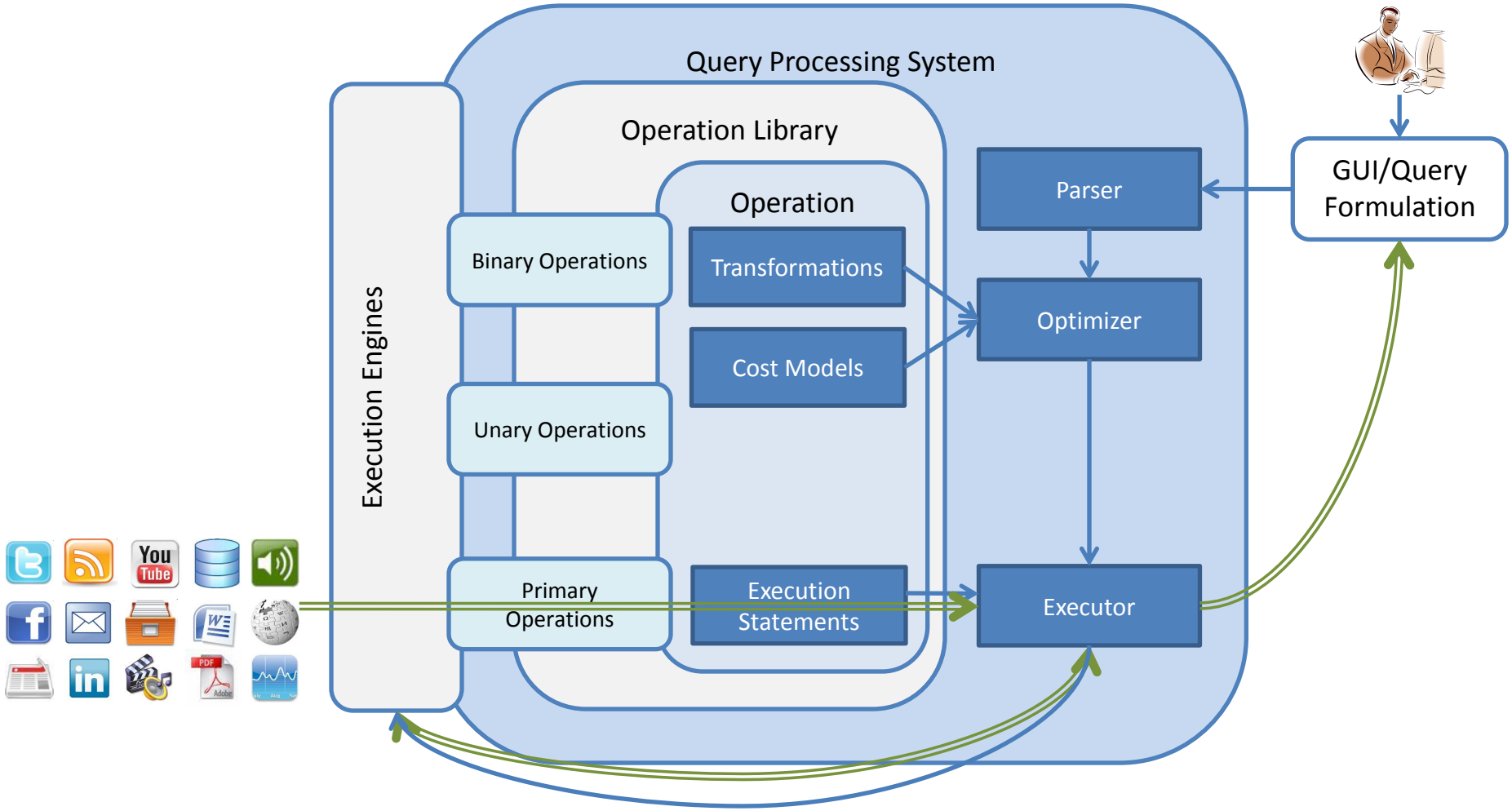
Список операций

- Первичные операции
  Извлечение информации из первичных источников
- Фильтры (простые и обогащающие)
  Обработка объектов по одному
- (Нечеткие) теоретико-множественные операции
  Аргументы из одного базового множества
- (Нечеткие) соединения
- (Нечеткие) операции агрегирования
- Расщепление
- Группирующие соединения
  Комбинация соединения и агрегирования

Расширяемость

- Внешние фильтры
- Пользовательские операции
- Пользовательские предикаты
- Пользовательские функции оценки
- Пользовательские группирующие функции

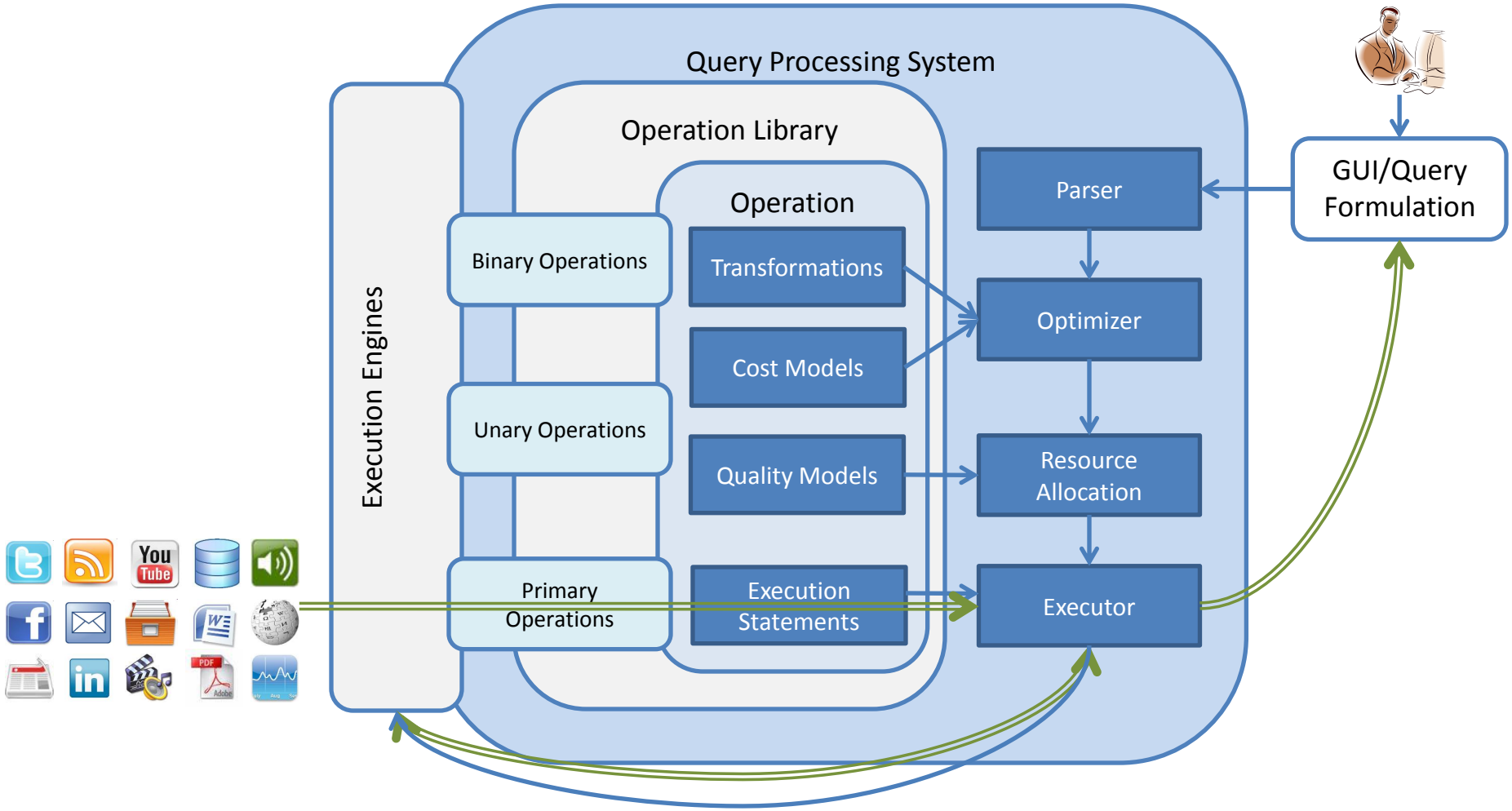
Архитектура с оптимизацией



Пересмотр задачи оптимизации

- Классическая (точное выполнение запросов): Найти план выполнения запроса с минимальной стоимостью
- Баланс между стоимостью и качеством
  - Гарантированное качество: Найти план обеспечиающий качество не ниже заданного за минимальное время
  - Ограниченная стоимость: Найти план обеспечивающий максимальное качество за ограниченное время
  - Многокритериальная оптимизация
- Эвристика: построить план и распределить ресурсы

Архитектура с распределением ресурсов



Приближенные алгоритмы

- С контролируемым качеством (выделенные ресурсы определяют качество результата)
- С фиксируемыми параметрами (фиксированное количество выделенных ресурсов трансформируется в фиксированные параметры вызова операции)

Качество

- Абсолютное качество
- Качество операции

  Отношение качество входа к качеству выхода
- Относительное качество

  Отношение абсолютного качества выхода при фиксированных ресурсах к абсолютному качеству при неограниченных ресурсах

Развитие модели стоимости

- Базовая (традиционные оптимизаторы)

  Resource = operation_cost(argument_size)

- Расширенная

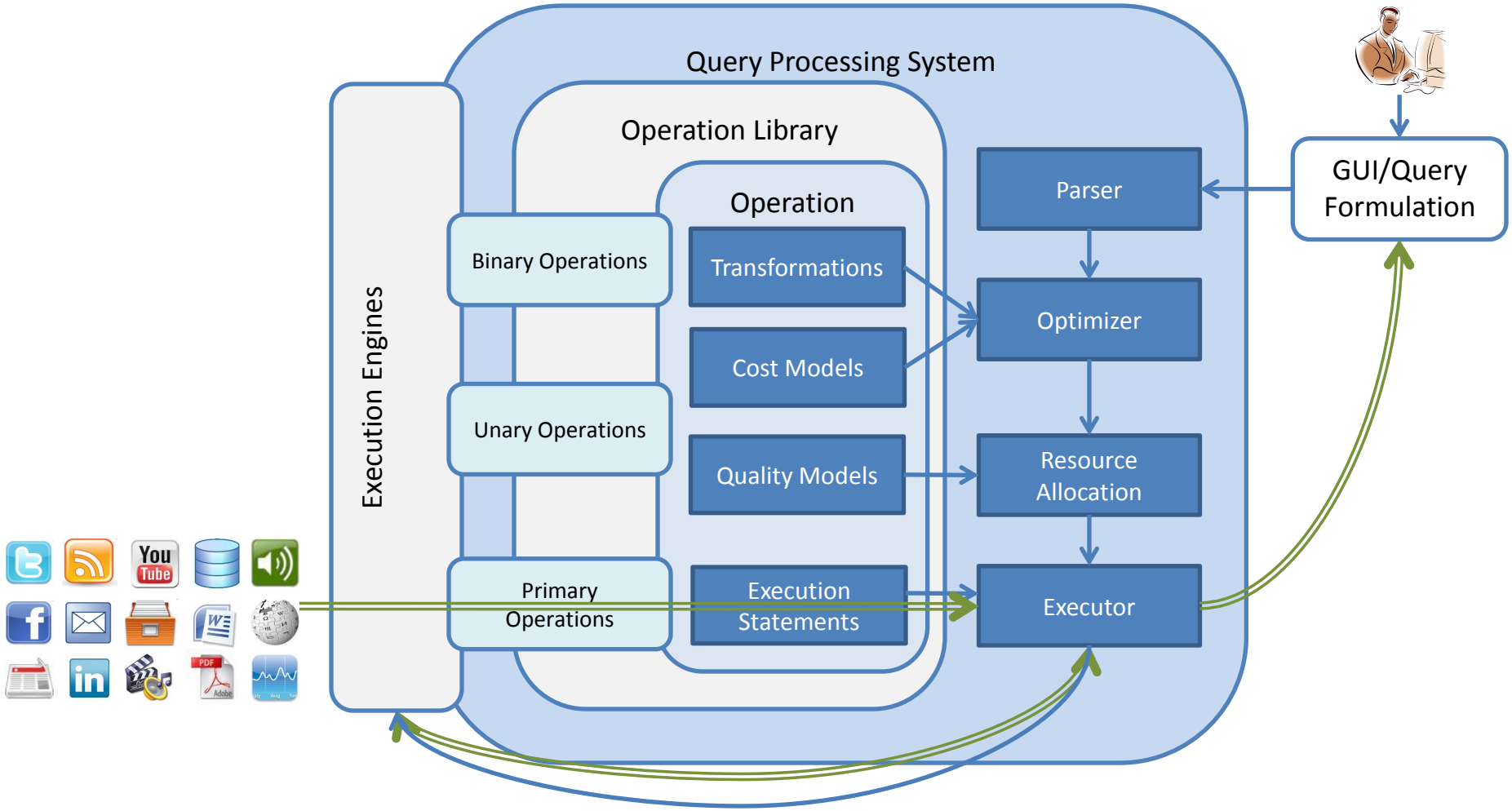
  Resource = operation_cost(argument_size, quality)

  Quality=operation_quality(argument_size, resource)

Модели качества

- Модель качества
  - Input: Статистики входных данных; конфигурация операции
  - Output: Кусочно-линейное представление зависимости между ресурсами, выделенными на выполнение операции, и качеством результата
- Функция отображения
  - Input: Статистики входных данных; конфигурация операции; количество ресурсов, выделенных на выполнение операции; ожидаемое качество результата
  - Output: Параметры вызова операции

Архитектура системы

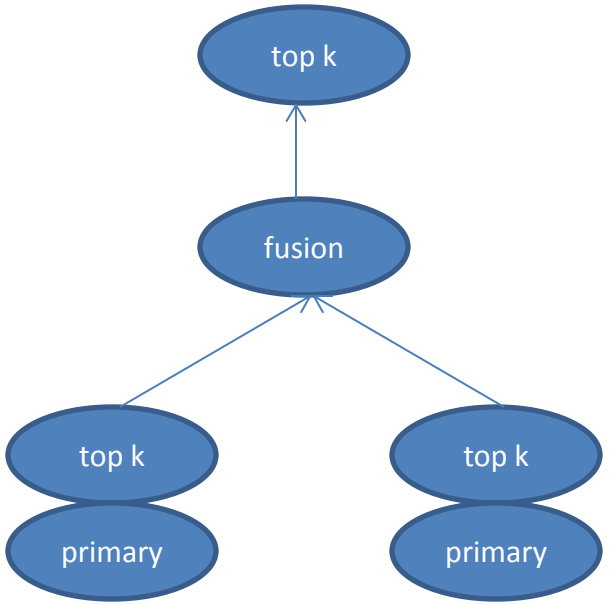


Задача распределения ресурсов

Для заданного плана выполнения запроса найти оптимальное распределение фиксированного ресурса, обеспечивающего максимальное качество

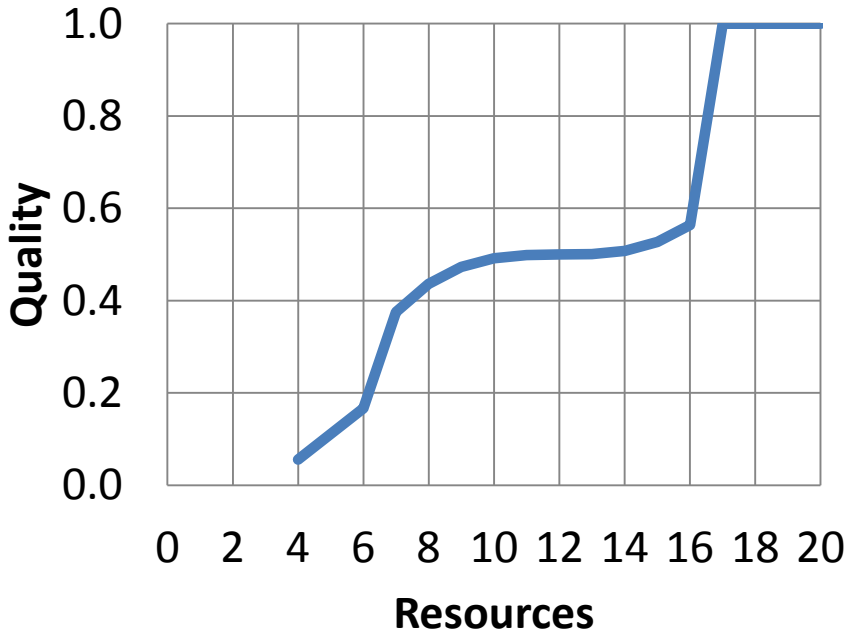
Используемые обозначения

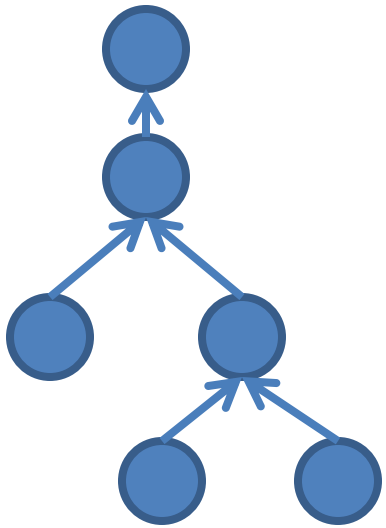
- Дерево плана P
- Операции (алгоритмы)
- Функции качества
- Ресурсы (время) T



Функции качества

- $q(x) : \mathcal{R} \to Q$ (операция)
- $Q(x): \mathcal{R} \to Q$(план)
- $Q(x) = \min(Q(l); Q(r))q(x)$



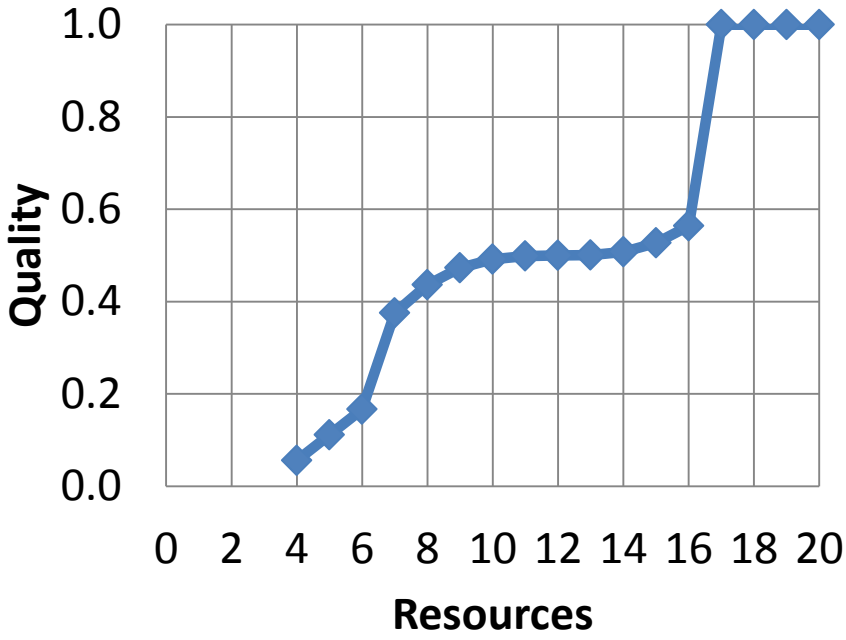


Задача оптимизации (распределения ресурсов)

- Дерево плана P и фиксированное количество ресурса T
- Для любой операции $x \in P$ выделить ресурс $t_x \in \mathcal{R}$ так, что $\sum t_x \leq T$ и $Q(root)$ максимально
- Множество $t_P = \{t_x : x \in P\}$ называется распределением ресурсов

Предположения

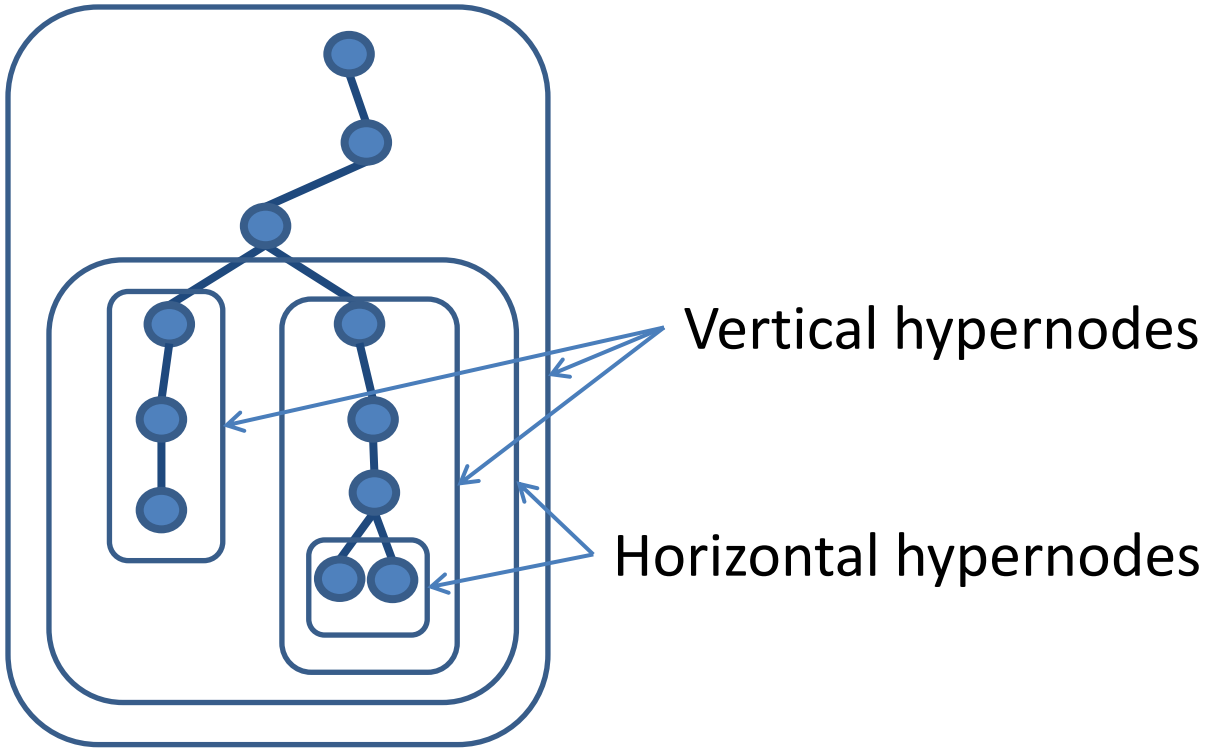
- Границы
  $t_{min}(x), t_{max}(x) \in \mathcal{R}$
- Качество выхода неубывающая функция от выделенного количества ресурса
  $\forall x \in P;\ t_1, t_2 \in \mathcal{R}:\ t_1 < t_2$
  $q(x)(t_1) < q(x)(t_2)$
- Качество операции может быть приближено непрерывной кусочно-линейной функцией



Подход к решению

- Леммы о распределении ресурсов в частных случаях:
  - Вдоль пути
  - Между братьями
- Алгоритм
  - Разбиение на фрагменты
  - Итеративное приближенное решение для произвольного плана

Гиперграф



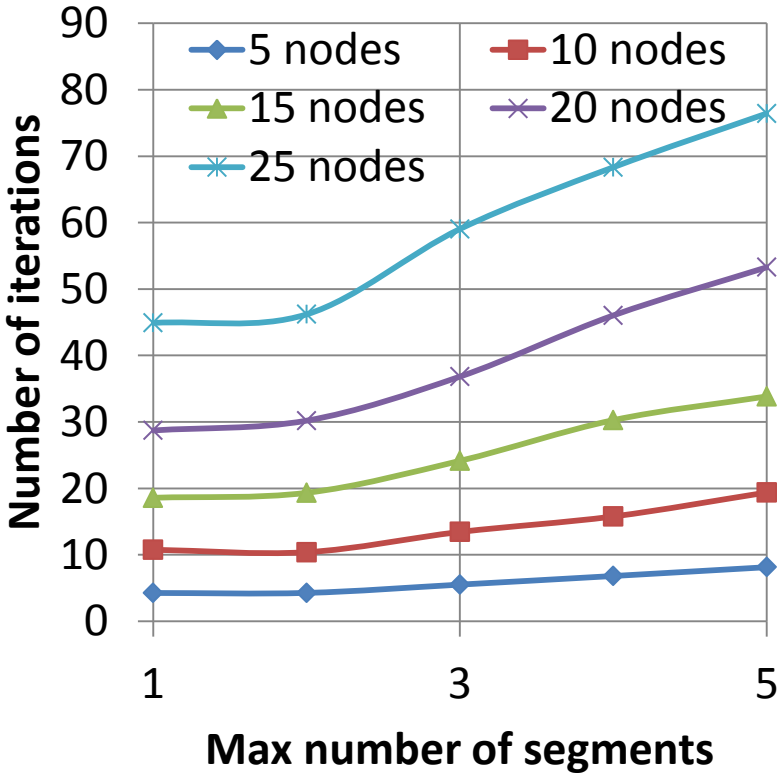
Алгоритм распределения ресурсов

- Input: Дерево плана P, ресурс T
- Output: Для всех операций $x \in P$ распределенное количество ресурса t_x

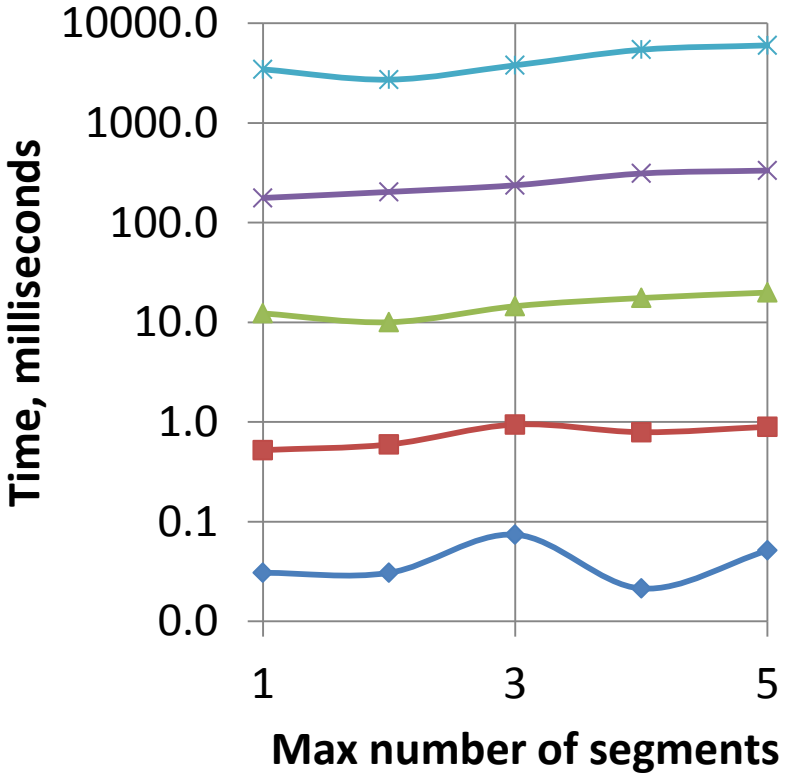
```
Initialization(P;T)
while T > 0 & !isMaxQualityReached(P) do
    H=HypergraphConstruction(C)
    ResourceAllocation(H;T)
    QualityEstimation(P)
end while
```

Производительность

Среднее число итераций

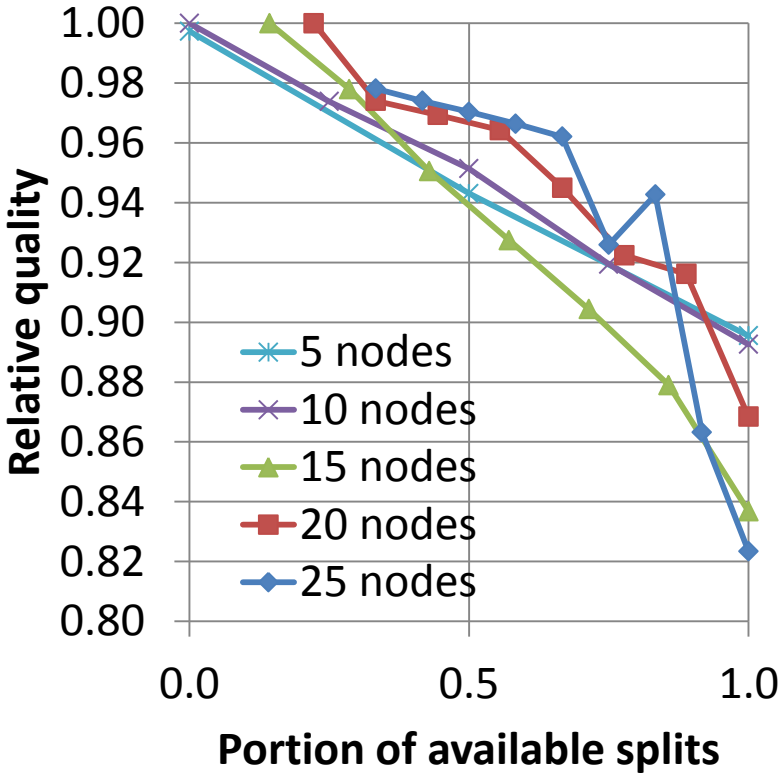


Абсолютное время

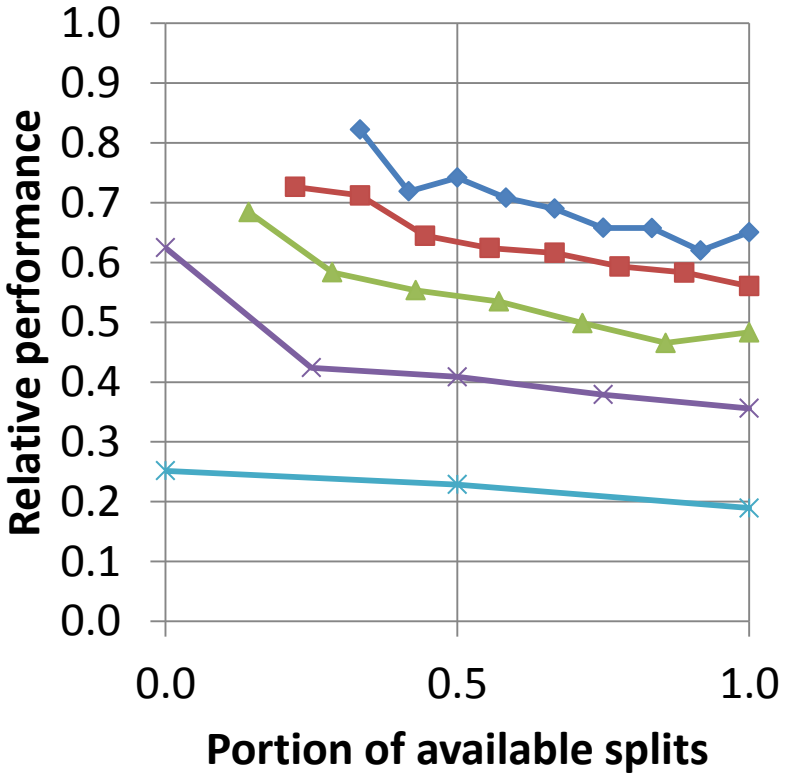


Баланс между стоимостью и качеством

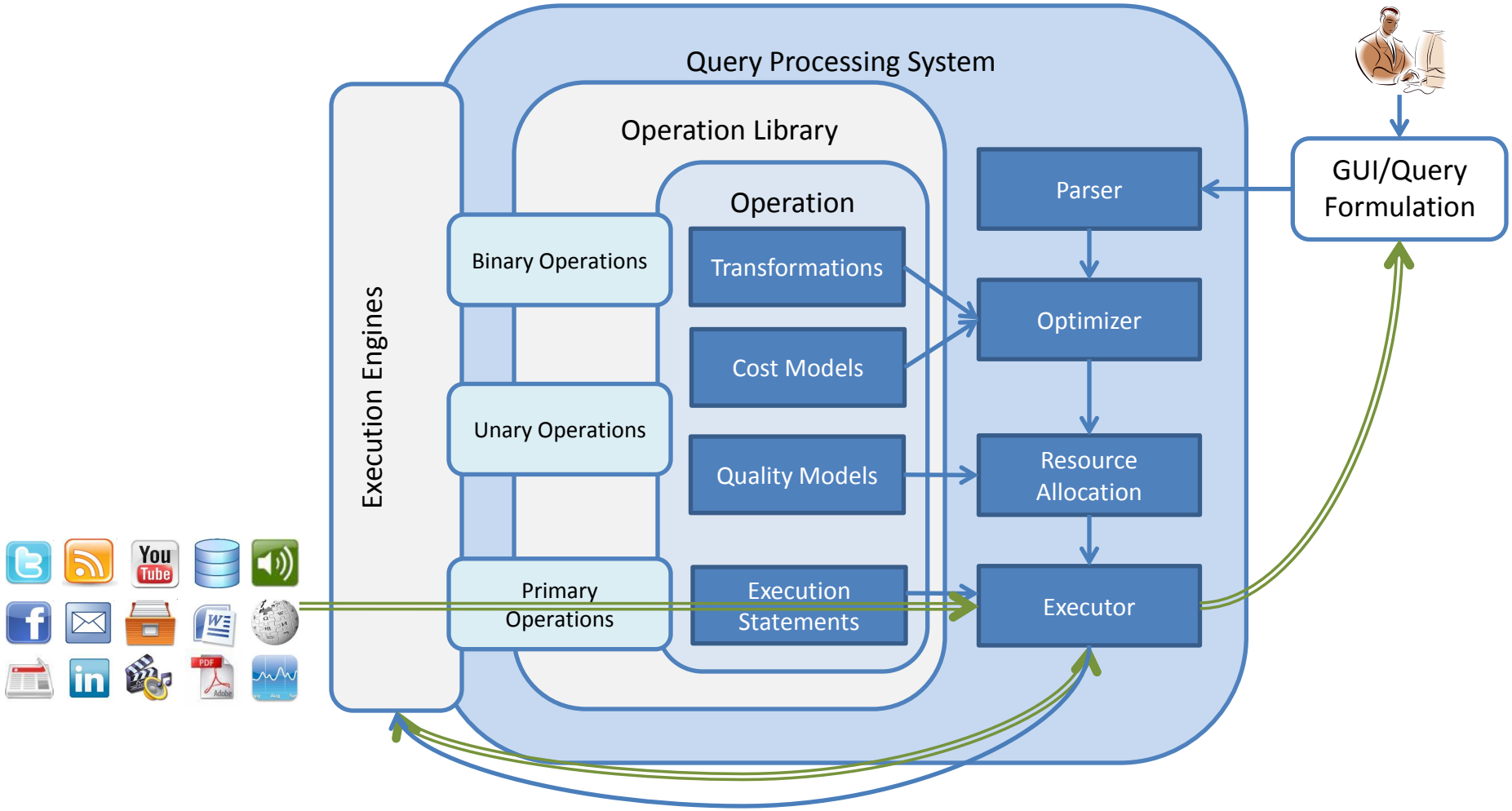
Относительное качество



Относительная производительность



Оптимизация и выполнение запросов



План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение

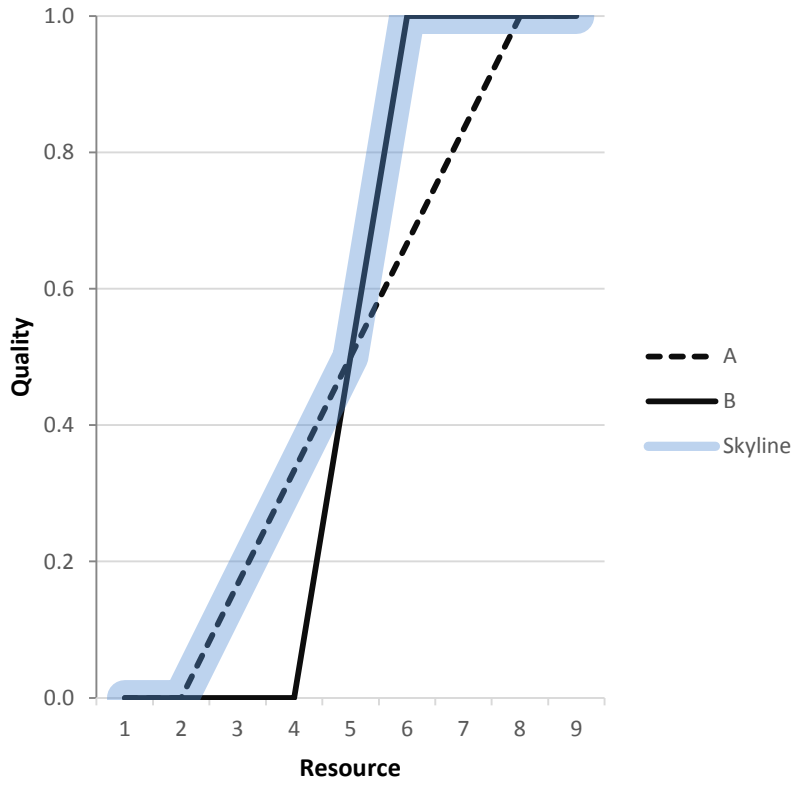
Сценарий выполнения запроса

- Формулирование запроса (и ограничений)
- Оптимизация запроса
  - Зависимости критериев
  - Множество оптимальных планов
- Исполнение
  - Уточнение ограничений
  - Оптимальный план
  - Конфигурация плана
  - Исполнение плана

Time	Quality
5 msec	10%
5 sec	30%
10 min	80%
24 hours	99%
1120 years	99.999%

Задача многокритериальной оптимизации

Многокритериальная оптимизация

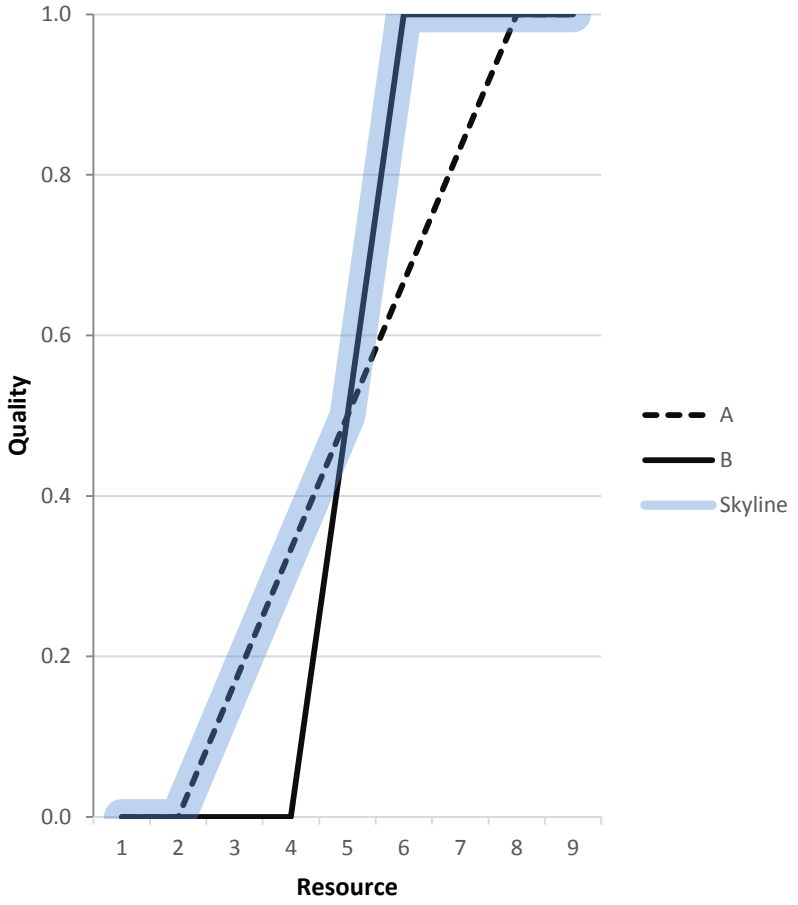


Подход

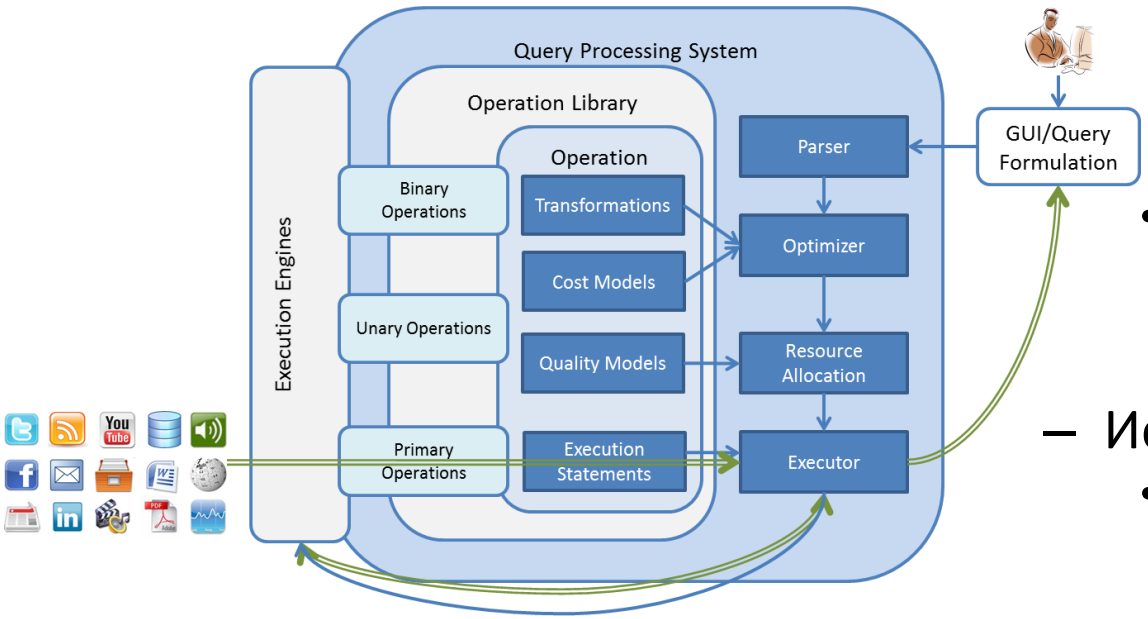
- Функция качества плана
  - Q_P: $\mathcal{R} \rightarrow Q$
  - Свойства
    - $t_{min}, t_{max} \in \mathcal{R}$
    - $\forall t_1, t_2 \in \mathcal{R}: t_1 < t_2$ $Q_P(t_1) < Q_P(t_2)$
- Функция качества стратегии
  - Q_S: $\mathcal{R} \rightarrow Q$
  - Сегменты ассоциированы с планом выполнения запроса

Доминанта

Доминанта запроса это функиця качества, доминирующая над функциями качества всех планов (стратегий) выполнения запроса



Решение задачи многокритериальной оптимизации



- Оптимизация
  - Построить функцию качества плана на основе моделей стоимости/качества и стратегии оптимального распределения ресурсов
  - Построить доминанту запроса на основе алгоритма перечисления планов
- Исполнение
  - Выбрать оптимальный план выполнения запроса для заданных ограничений на основе доминанты

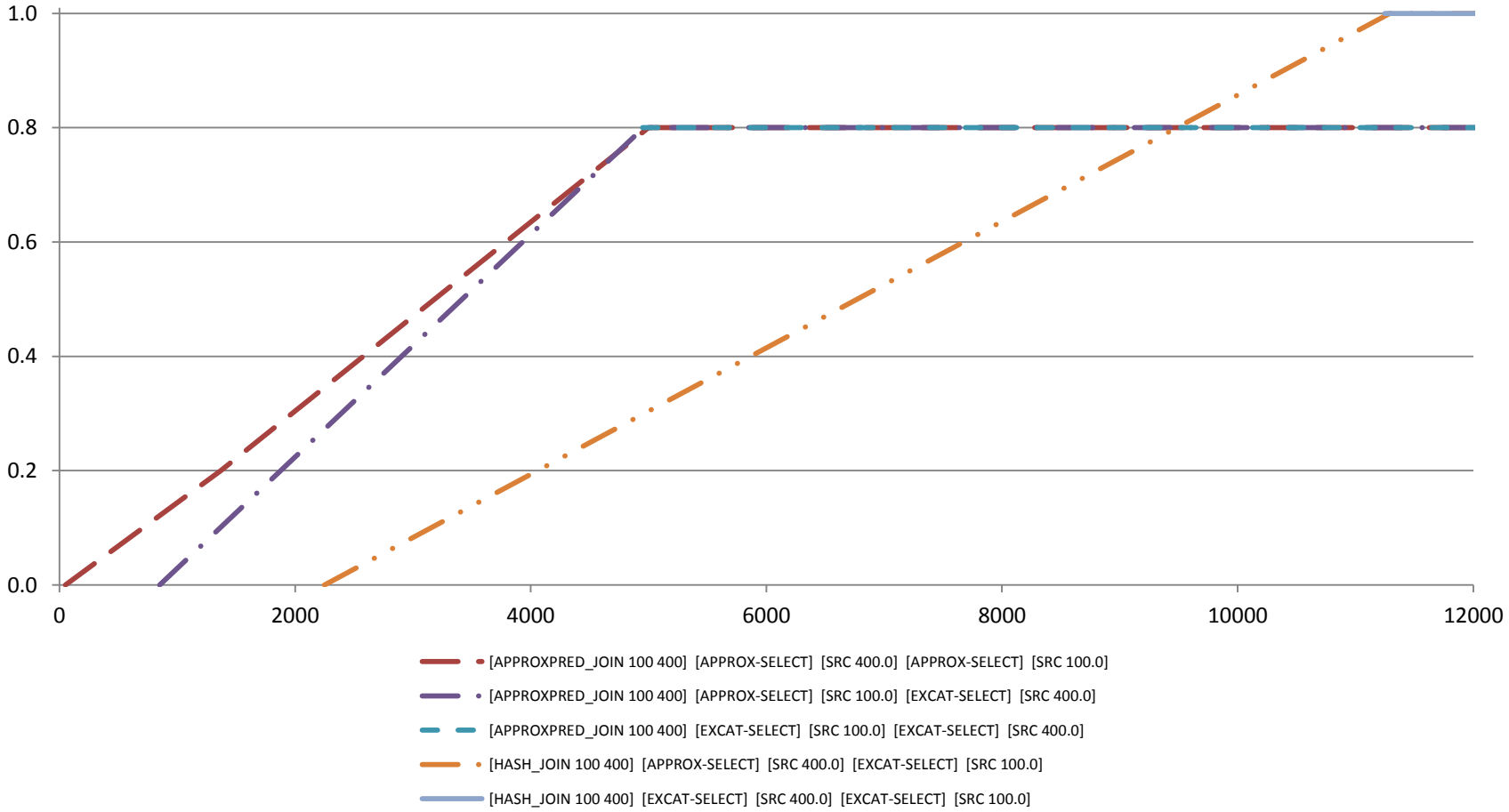
Доминанта для запроса

- Слияние функций качества
- Построение доминанта для запроса (перечисление планов)
  - Полный перебор
  - Снизу-вверх
  - Сверху-вниз
- Сжатие функций качества

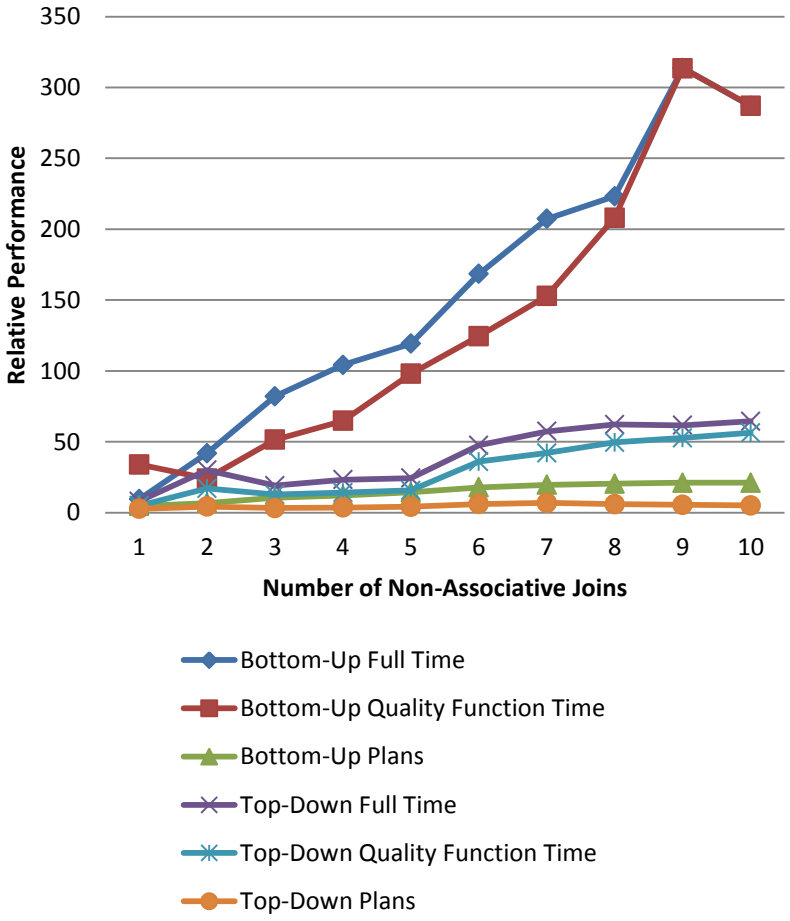
Цели экспериментов

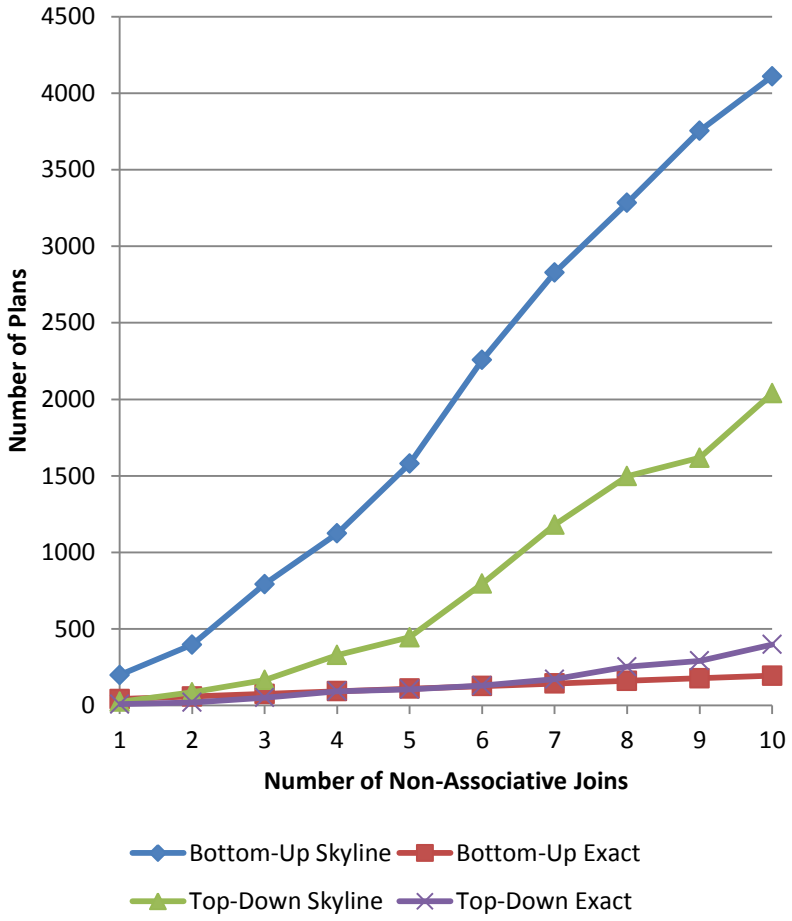
- Существование нетривиальной доминанты для типичных запросов;
- Производительность построения доминанты запроса
- Точность приближенных доминант

Нетривиальная доминанта



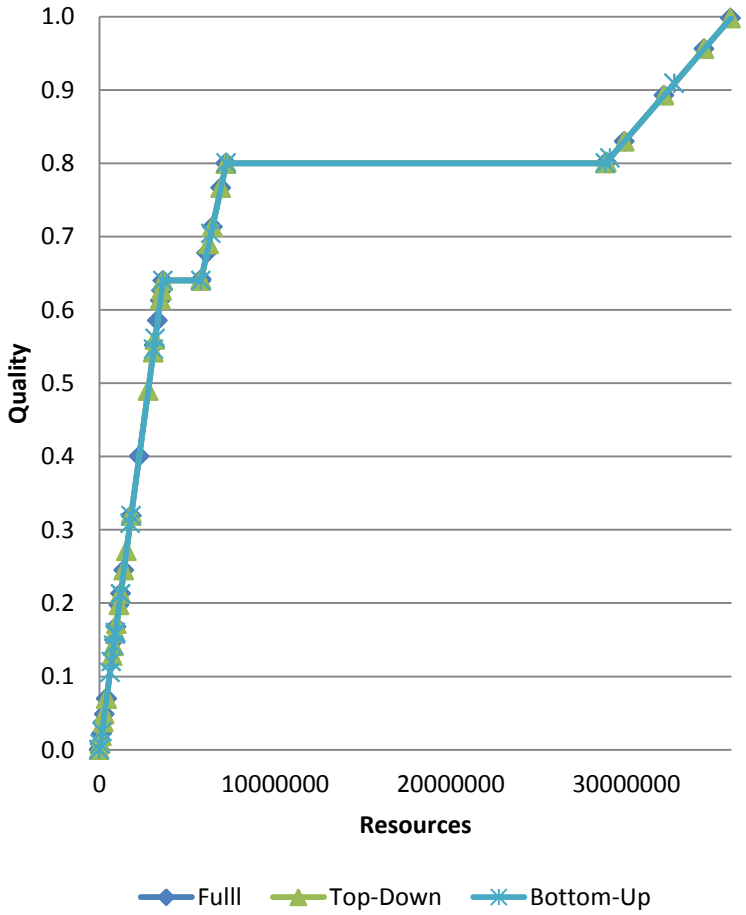
Относительная производительность построения доминант



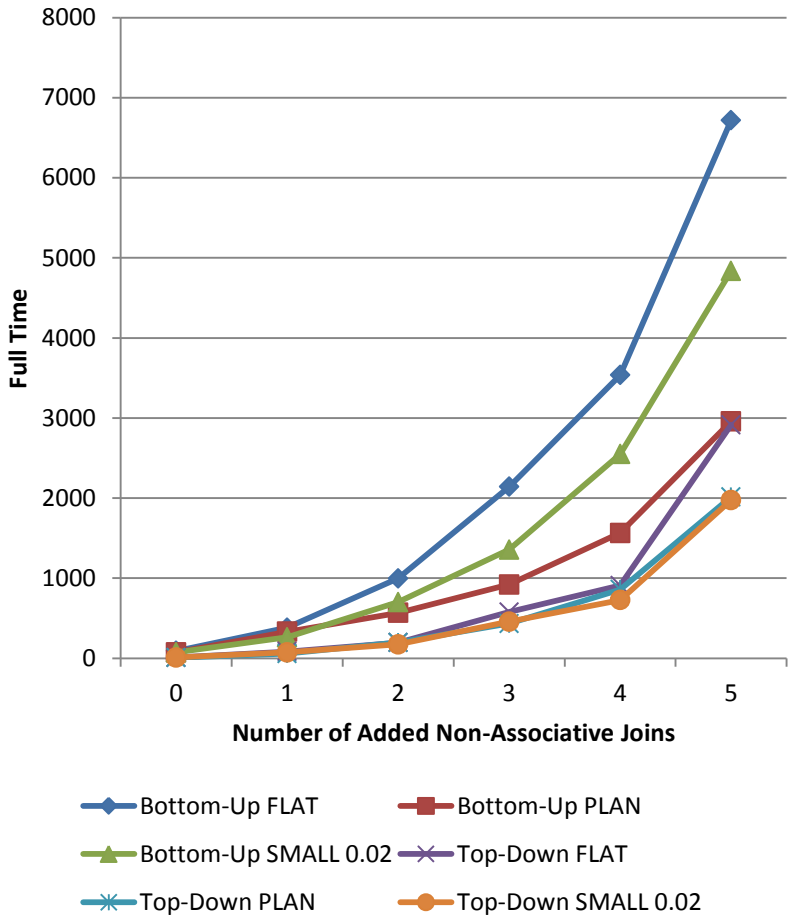


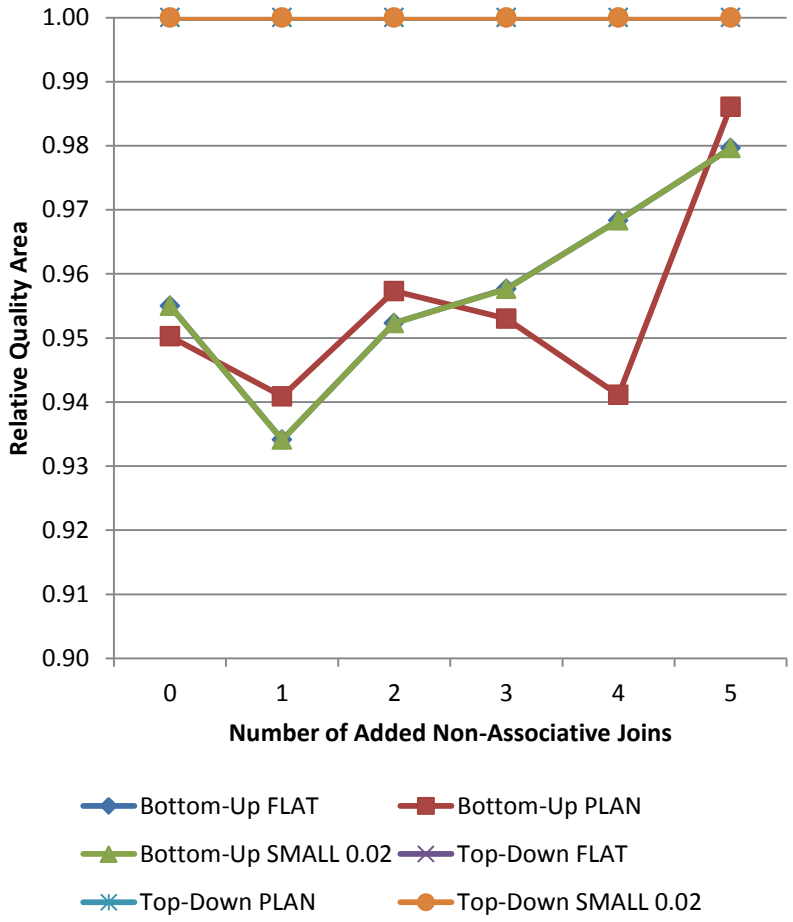
Построение доминант простых запросов

Algorithm	Full Time	Skyline Time	Visited Plans
SSSJJ			
FULL	562	271	576
BOTTOM_UP	246	119	441
TOP_DOWN	171	56	259
SSSSJJJ			
FULL	36719	22673	17280
BOTTOM_UP	2623	1428	2952
TOP_DOWN	752	325	911
SSFFJ			
FULL	238	106	288
BOTTOM_UP	94	38	226
TOP_DOWN	21	11	44



Построение доминант со сжатием





Заключение

- Расширяемая алгебра операций
- Расширенная модель стоимости (качества)
- Алгоритм распределения ресурсов
- Алгоритмы многокритериальной оптимизации

План доклада

- Введение
- Обзор литературы
  - Выполнение точных запросов
  - Адаптивное выполнение запросов
  - Выполнение приближенных запросов
  - Многокритериальная и параметрическая оптимизация
- Подходы к решению и основные результаты
  - Теоретическая модель оптимизации и контролируемого приближенного выполнения нечетких запросов
  - Архитектура системы приближенного выполнения нечетких запросов в реальном времени
  - Задача распределения ресурсов
  - Многокритериальная оптимизации запросов при специфических ограничениях
- Заключение