Дата последней сборки документа:
4 августа 2020 г. в 20:49

Заметки по большим данным и облачным технологиям

Подвойский А.О.

Здесь приводятся заметки по некоторым вопросам, касающимся больших данных, облачных технологий, машинного обучения, анализа данных, программирования на языках `Python`, `R` и прочим сопряженным вопросам так или иначе, затрагивающим работу с данными.

Содержание

1. Основные термины и определения

Витрина данных (Data Mart) – срез хранилища данных, представляющий собой массив тематической, узконаправленной информации, ориентированный, например, на пользователей одной рабочей группы или департамента.

2. Логическая витрина для доступа к большим данным

Пример. Рассмотрим некоторый промышленный комплекс, обладающий огромным количеством оборудования, обвешанного различными датчиками, регулярно сообщающими сведения о состоянии этого оборудования. Для простоты рассмотрим только два аггрегата (котел и резервуар), и три датчика (температуры котла и резервуара, а также давления в котле).

Эти датчики контролируются АСУ разных производителей и выдают информацию в разные хранилища: сведения о температуре и давлении в котле поступают в HBase, а данные о температуре в резервуаре пишутся в лог-файлы, расположенные в HDFS.

Данные о датчиках могут храниться, например, в `PostgreSQL`, а показания этих датчиков – в HDFS, HBase и т.п. Теперь пусть мы хотим предоставить аналитику возможность делать запросы. Заранее построить и запрограммировать сложные запросы не получится. Выполнение любого сложного, тяжелого запроса требует связывания данных из разных источников, в том числе из находящихся за пределами нашего модельного примера. Извне могут поступать, например, справочные сведения о рабочих диапазонах температуры и давления для разных видов оборудования, фасетные классификаторы, позволяющие определить, какое оборудование является маслонаполненным и др. Все подобные запросы аналитик формулирует в терминах концептуальной модели предметной области, то есть ровно в тех выражениях, в которых он думает о работе своего предприятия.

Витрина данных – предметно-ориентированная и, как правило, содержащая данные по одному из направлений деятельности компании база данных. Она отвечает тем же требованиям, что и хранилище данных, но в отличие от него, нейтрально к приложениям. В витрине информация храниться оптимизированно с точки зрения решения конкретных задач.

Витрины данных имеют следующие достоинства:

- пользователи ведят и работают только с теми данными, которые им действительно нужны,
- для витрин данных не требуется использовать мощные вычислительные средства.

К недостаткам витрин данных можно отнести сложность контроля целостности и противоречивости данных.

Список иллюстраций

Список литературы

1. *Лутц М.* Изучаем Python, 4-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2011. – 1280 с.