

Pentaho Enterprise: интеграция и аналитика по модели самообслуживания

Self-Service Data Preparation для экосистемы больших данных и не только

Александр Тимчур

Hitachi Vantara | Analytic Solutions

Alexander.Timchur@HitachiVantara.com

Платформа Pentaho Enterprise

- Один из лидеров рынка интеграции и аналитики больших данных
 - 12 лет на рынке аналитики;
 - Лидер в области интеграции больших данных последние 8 лет, создатели термина «Data Lake»;
 - Проверенное решение, работающее в масштабе петабайтов данных;
 - Наиболее полная поддержка технологий Big Data на рынке и интеграция с облачными решениями (Google / Amazon / Microsoft);
 - Единственная целостная платформа «полного цикла» от интеграции до машинного обучения и визуализации.

- **1500+ заказчиков в Enterprise** сегменте в 185 странах мира во всех отраслях

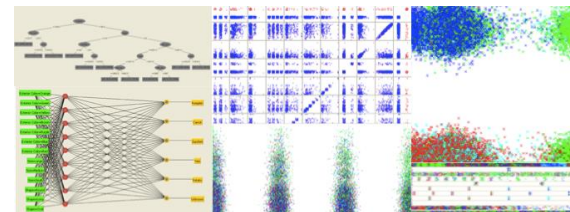
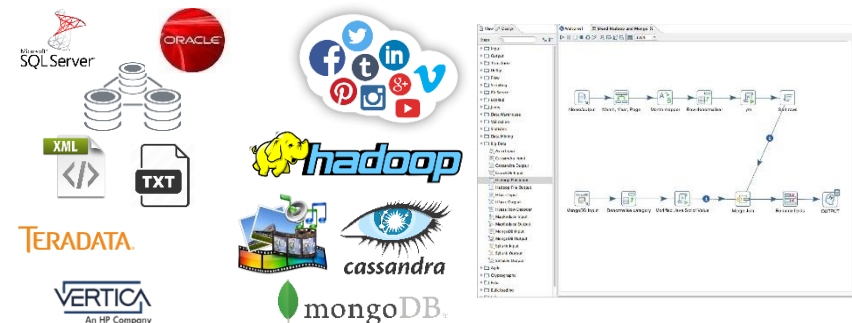


Глобальные и технологические партнёры



Единая аналитическая платформа

HITACHI
Inspire the Next



Интеграция данных и Big Data

- ✓ Традиционный ETL
- ✓ Интеграция структурированных и неструктурированных данных
- ✓ Очистка и обогащение
- ✓ Работа с NoSQL базами данных и Hadoop
- ✓ ELT Pushdown и оркестрирование Spark

Визуализация и анализ

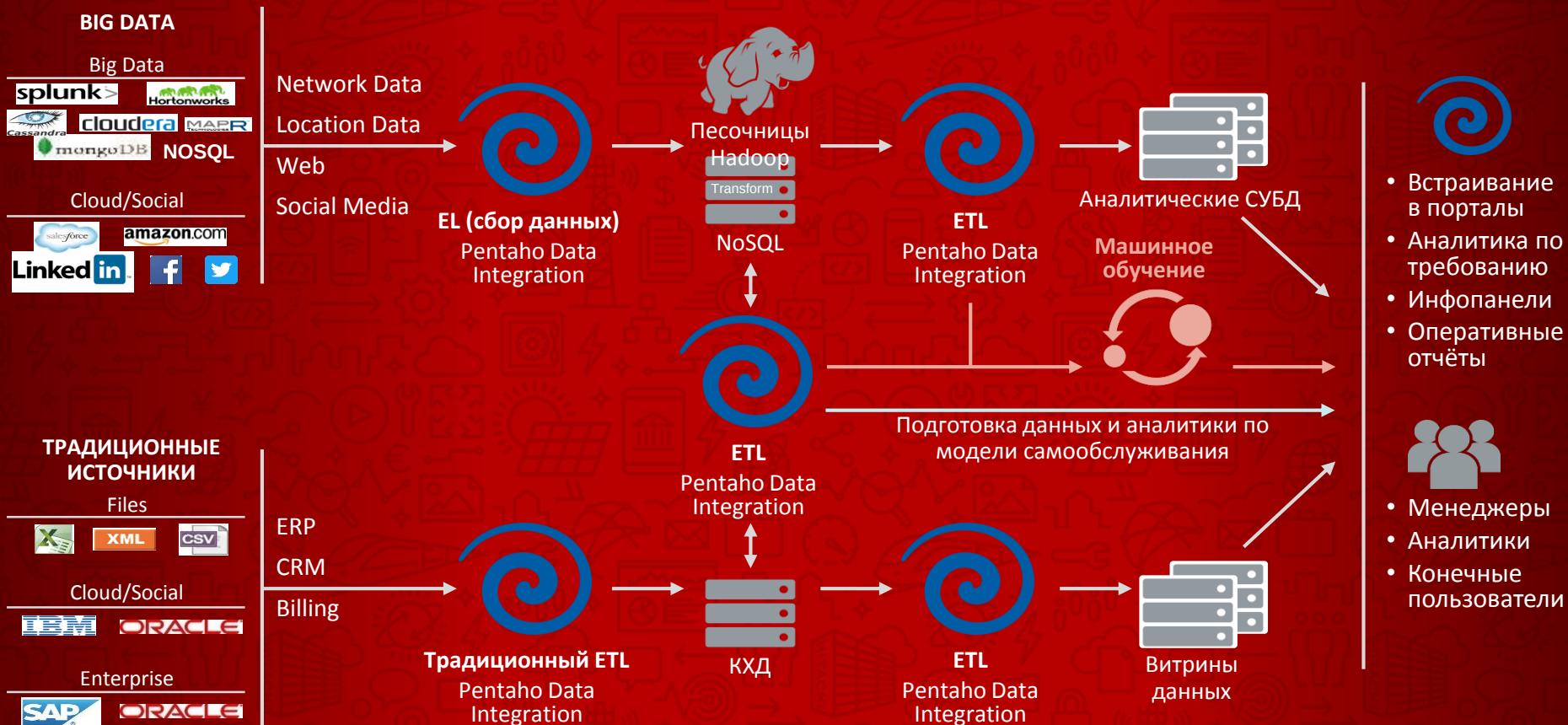
- ✓ Визуализация и OLAP-анализ
- ✓ Построение инфопанелей и отчётности
- ✓ Встраивание аналитики, API

Моделирование и Прогнозирование

- ✓ Решение задач продвинутой аналитики и Data Mining
- ✓ Интеграция с R, Python
- ✓ Включение аналитических блоков в поток обработки данных (операционная аналитика)

Возможности применения Pentaho

HITACHI
Inspire the Next



Возможность подключения ко всем типам источников данных

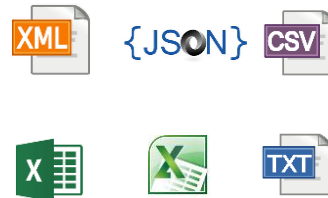
Hadoop + SQL/H



MPP / АСУБД



Файлы



NoSQL



Традиционные БД



Интерфейсы



salesforce.com
Success On Demand™

- Большинство заказчиков работает больше чем с одним типом СУБД
- Как правило используется Hadoop + SQL on Hadoop
- Файловые источники никто не отменял (XML, Excel, Json)
- Различные программные интерфейсы для разработки и управления
- Вершина пирамиды: SAP системы

Self-Service Data Preparation, а не Self-Service Analytics

Никто не сделает витрины данных за вас, когда
у IT нет свободных ресурсов

Предпосылки появления практики подготовки данных и аналитики по модели самообслуживания

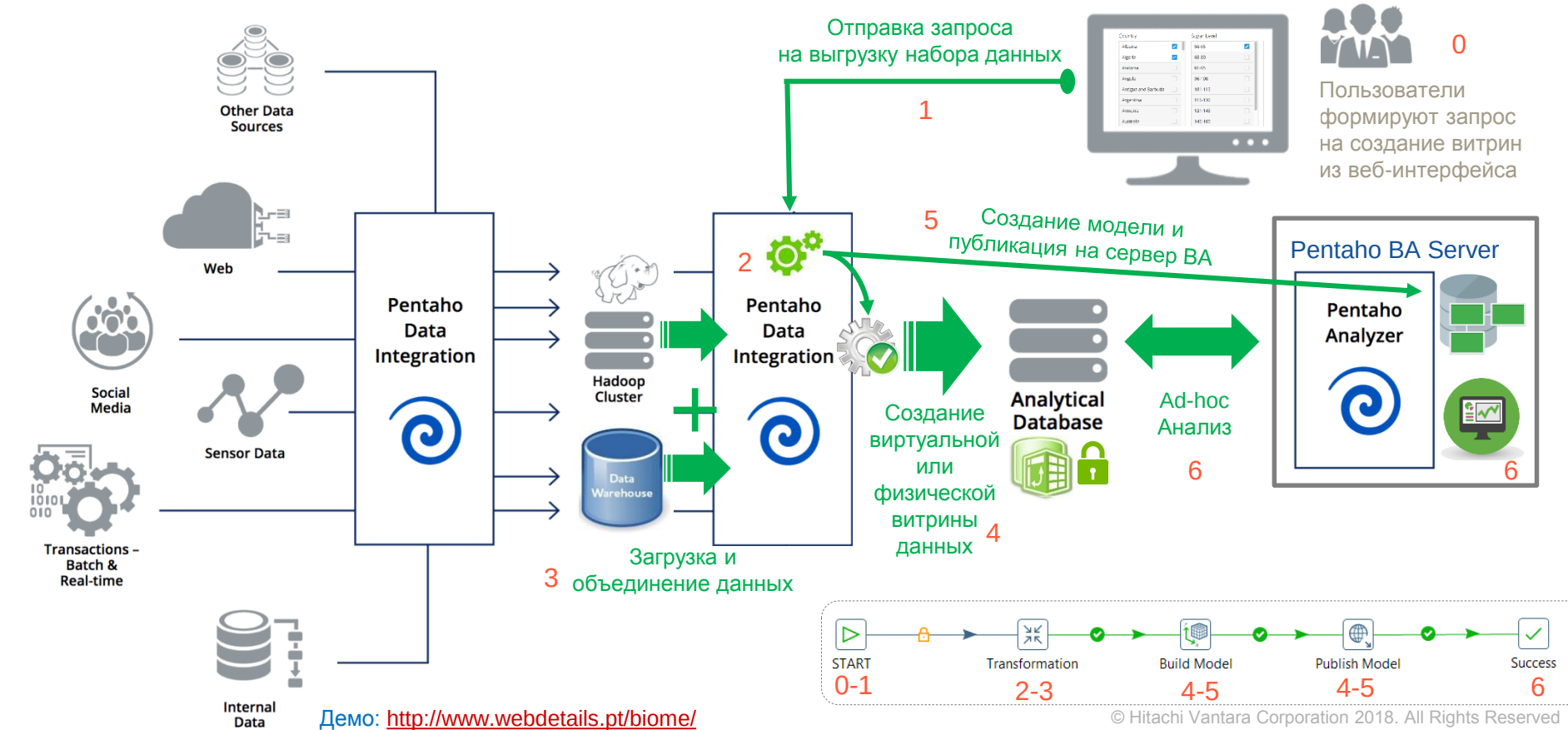
- **Time to market** (“анализ / отчёт нужен позавчера”);
- Периодические **изменения форматов** источниках данных (*не управляется*);
- **Сокращение бюджетов** на полный цикл подготовки аналитики (*BRD / FSD / разработка / тестирование / внедрение*);
 - Сокращение внутреннего IT → аутсорсинг;
 - Сокращение затрат на подрядчиков;
- Необходимость дать **доступ аналитикам к исходным данным** (*raw data*);
- Аналитикам **необходимо подготавливать данные для себя**, но доступа к Enterprise ETL нет (*и не будет*);
- **Data Science практика** (*ближе к данным как только можно*);

>> [Self-Service Data Preparation <> Self-Service Analytics](#)



Self-Service Data Preparation для бизнес-пользователей

Работа по модели самообслуживания



Демо: <http://www.webdetails.pt/biome/>

Self-Service Data Preparation для бизнес-пользователей

С точки зрения конечного пользователя

Avg. Heart Rate

All Requests Processed

Country

Sugar Level

Daily Calories Burned

Gender

Age Group

Years

Albania

Algeria

Andorra

Angola

Antigua and Barbuda

Argentina

Armenia

Australia

50-65

66-80

81-95

96-100

101-115

116-130

131-145

146-160

0-500

501-1000

1001-1500

1501-2000

2001-2500

2501-3000

3001-3500

3501-4000

Female

Male

0 to 10

11 to 20

21 to 30

31 to 40

41 to 50

51 to 60

61 to 70

71 to 80

2005

2006

2007

2008

2009

2010

2011

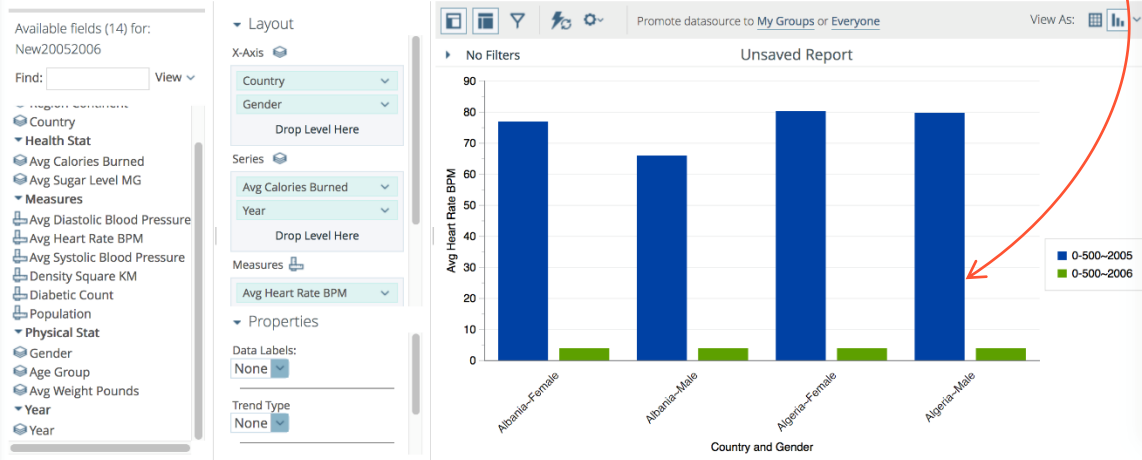
2012

Permission

Data Source Name

Analyze

Шаг 1
Отбор необходимых разрезов пользователем для отправки на ETL-процессинг



Шаг 2
Получение результата и дальнейший анализ

Пользовательский ETL

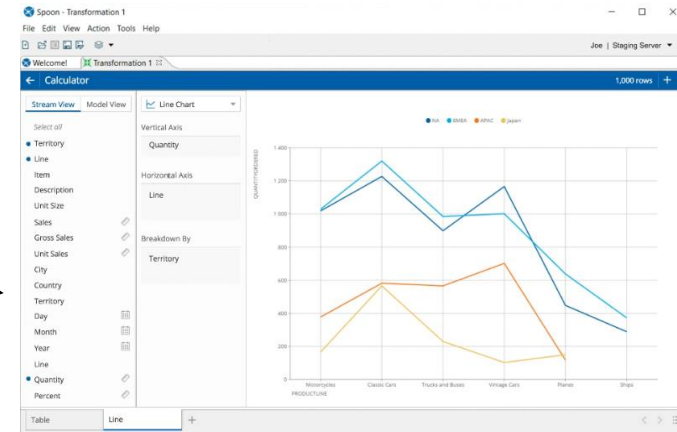
для аналитиков с активной жизненной позицией

Обученные аналитики используют мост между DI и BA компонентами и создают собственные витрины и наборы данных с проверкой на каждом шаге трансформации – визуально.

The screenshot shows the Spoon - Transformation 1 interface. The main window displays a data flow from 'CSV File Input' to 'Table output' via a 'Select values' component. A 'Calculator' component is highlighted with a dashed circle, and an arrow points to a detailed view of the Calculator component. The Calculator component is shown in 'Table' view, displaying a list of data rows with columns for Territory, Line, Item, Description, Unit, Sales, Gross, and Unit Sales. A 'Select a visualization type' dialog is open, showing various chart types like Bar, Line, and Pie.

The screenshot shows the Spoon - Transformation 1 interface with a Pivot Table view. The Pivot Table is configured with 'Territory' as the Row, 'Projected' as the Column, and 'Years' (2003, 2004) as the Values. The data is displayed in a table format.

Territory	Projected	Projected
APAC	\$19,325	\$24,458
EMEA	\$20,084	\$10,002
Japan	\$10,325	\$11,458
NA	\$30,084	\$50,002

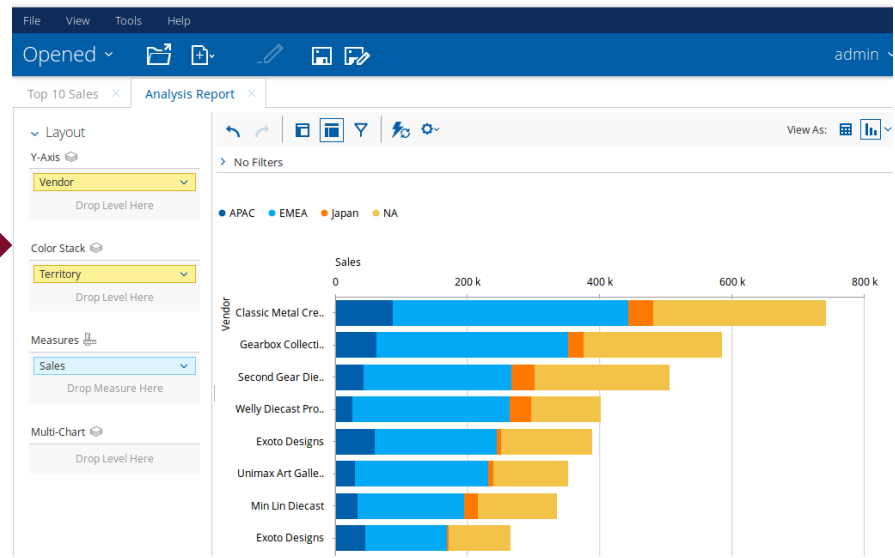


Публикация виртуальных витрин для анализа

Опция для аналитиков без прав создания витрин



T...	Line	Item	Description	Un...	Sales	Gros...	Unit...	Sale...
APAC	Feed	Alfalfa	50lb Alfalfa square bale	25.458	154.154	54.124	25.458	154.154
APAC	Feed	Bahia	50lb Bahia square bale	23.145	187.654	98.542	23.145	187.654
APAC	Feed	Bermuda	50lb Bermuda square bale	22.447	154.963	65.487	22.447	154.963
APAC	Feed	Coastal	50lb Coastal square bale	26.654	154.154	95.324	26.654	154.154
APAC	Feed	Orchard	50lb Orchard square bale	23.145	187.654	65.347	23.145	187.654
APAC	Feed	Timothy	50lb Timothy square bale	22.447	154.963	54.124	22.447	154.963
APAC	Feed	Wheat	40lb bag whole wheat	12.985	154.154	98.542	12.985	154.154
APAC	Feed	Oats	40lb bag whole oats	13.457	187.654	65.487	13.457	187.654
APAC	Feed	Mixed Grains	40lb bag corn, oats, sunfl...	11.674	154.963	95.324	11.674	154.963
APAC	Feed	Laver Pellets	20lb Purina Laveria	31.145	154.154	65.347	31.145	154.154
APAC	Feed	Grower Pellets	20lb Purina Flock Raiser	21.145	187.654	54.124	21.145	187.654
APAC	Feed	Starter Pellets	20lb Purina Fock Starter	22.124	154.963	98.542	22.124	154.963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154.154	65.487	451	154.154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31.458	154.963	65.487	12.124	154.963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154.154	65.487	451	154.154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31.458	154.963	65.487	12.124	154.963
APAC	Hardware	1/4" Hardware cloth	6ft x 100ft welded wire	451	154.154	65.487	451	154.154
APAC	Hardware	4" Hog Panel	4 foot hog panel	31.458	154.963	65.487	954	154.963



Agile аналитика:

Возможность подготовить собственную виртуальную витрину данных и опубликовать ее для коллег - без физической записи данных в DWH / КХД.

Self-Service Data Preparation для бизнес-пользователей

Безопасный процесс, удобный для всех

HITACHI
Inspire the Next



1. **Бизнес-пользователи** получают доступ к необходимым данным любого объёма в источниках необходимых систем с минимальной поддержкой со стороны ИТ
2. **Бизнес-пользователи** самостоятельно (Agile!) создают и проверяют аналитические витрины – физические или виртуальные - за счёт поддержки модели самообслуживания
3. **Оперативное получение пользователями необходимых для анализа данных** за счёт высокопроизводительного механизма обработки данных Pentaho Data Integration
4. **Автоматическая публикация моделей данных** для проведения дальнейшего анализа и визуализации
5. **Полнофункциональный удобный модуль OLAP-визуализации** Pentaho Analyzer
6. Возможность подключения (Python / R) или создания собственных **моделей прогнозной аналитики**

Работа с SAP ERP и SAP BW по модели самообслуживания



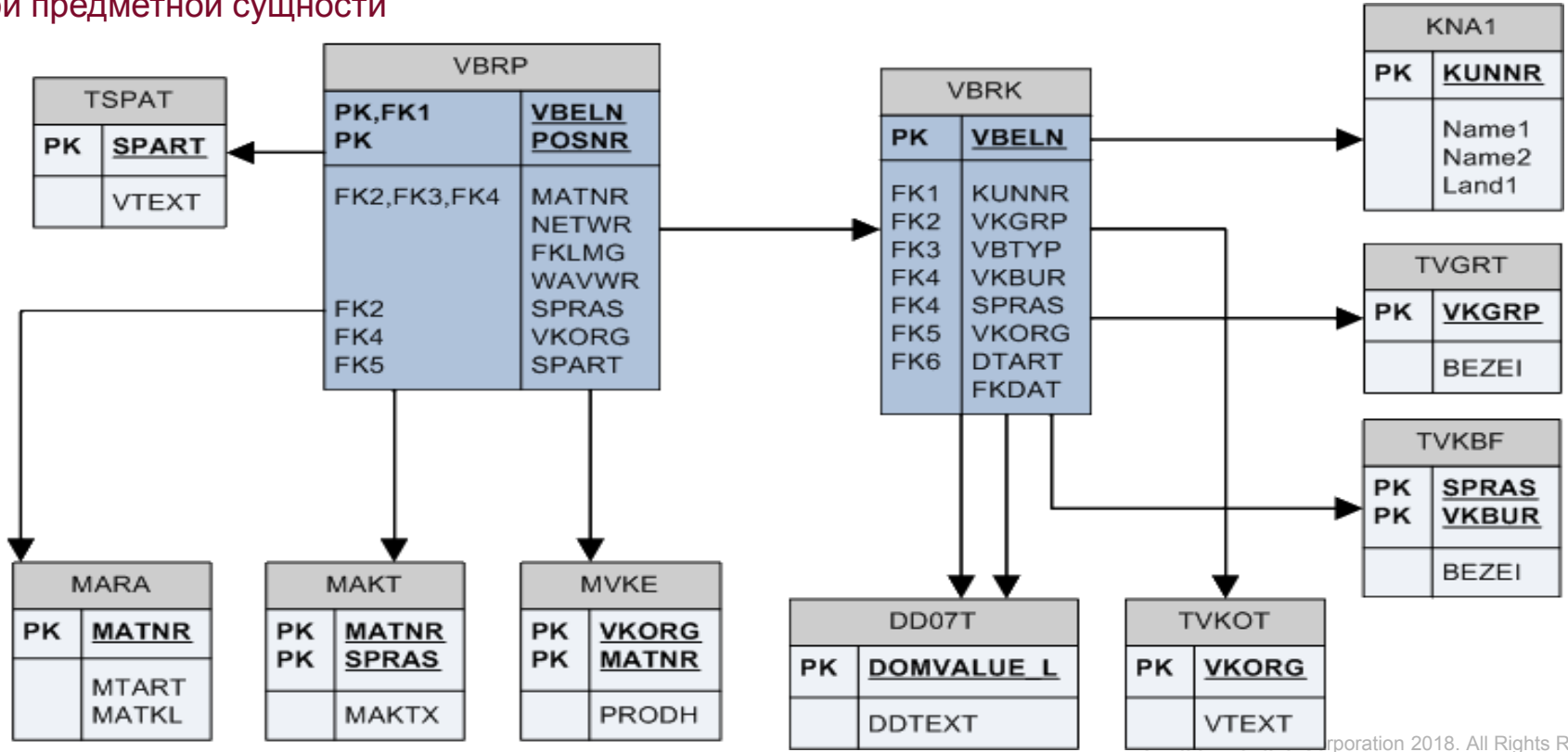
- 74% всех транзакций в мире так или иначе проходит через SAP-системы;
- SAP обслуживает более 300К заказчиков в 190 странах.. ..и 1500+ заказчиков в СНГ;
- Внешняя интеграция с SAP проблематична:
 - данные хранятся в сложных структурах;
 - доступ нетривиален и совершенно недоступен обычным аналитикам;
 - SAP-разработчики – бутылочное горлышко
(создание интерфейсов инфообмена или выгрузок данных).

Иными словами, SAP-приложения есть почти в каждом крупном заказчике, и у этих компаний есть задачи по интеграции или аналитике данных SAP систем и остального мира.



Пример: данные по продажам SAP R3 на уровне базы данных

Неясные акронимы таблиц и атрибутов, множество таблиц для хранения составляющих каждой предметной сущности



Решение проблемы интеграции: коннектор для SAP-систем

Расширение Pentaho для SAP от IT-Novum связывает инструменты Pentaho с источниками данных приложений SAP и позволяет выполнять интеграцию с системами вне контура SAP, расширенную аналитику или анализ обогащённых SAP-данных по модели самообслуживания.

Коннектор позволяет обрабатывать данные SAP ERP и SAP BW при помощи стандартных интерфейсов платформы Pentaho – визуально и без программирования.

Примеры использования



Пример: данные по продажам на выходе из IT-Novum SAP Connector в контексте предметной области

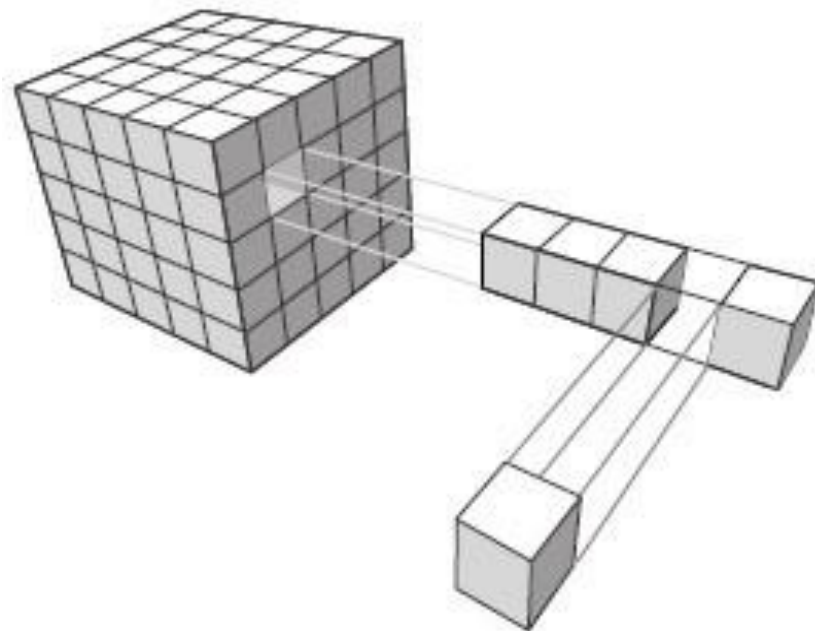
Фактические значения

- **Revenue** (VBRP.NETWR)
- **Units** (VBRP.FKLMG)
- **Invoice value** (VBRP.WAVWR)

Измерения

- **Data format - Actual, Plan** (VBRK.DTART)
- **Customer** (VBRK.KUNNR)
- **Material** (VBRP.MATNR)
- **Sales Office and Vendor** (VBRK.VKBUR, VKGRP)
- **Sales Receipt Type – Invoice, Cancellation Invoice** (VBRK.VBTYP)
- **Period** (VBRK.FKDAT)
- И многое другое...

**Sales Receipt, Invoice Position, Invoice Header,
Vendor Group, Sales Receipt Number, Position
Number, Invoice Date**



Q&A

HITACHI
Inspire the Next 