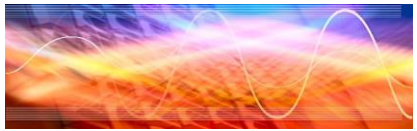
Big Data and Oracle Tools Integration: Kafka, Cassandra and Spark Creating Real Time Solutions

Dan Vlamis, President, Vlamis Software Solutions, Inc.
Jeff Shauer, Owner, JS Business Intelligence

Dan Vlamis

- Founded Vlamis Software Solutions in 1992
- 30+ years in business intelligence, dimensional modeling
- Oracle ACE Director
- BIWA Board Member since 2008
- BA Computer Science Brown University

Jeff Shauer



- Founder and owner of JS Business Intelligence
- Worked as a consultant, contractor, and employee
- Experience in developing, maintaining, owning, and managing financial planning systems at small, medium, and Fortune 500 corporations
- BIWA Board Member

Agenda

- Overview
- When to use Kafka, Spark, and Cassandra
- How to integrate Kafka, Spark, and Cassandra
- How to integrate Kafka, Spark and Cassandra into OBIEE / Tableau
- Tips and Tricks for Big Data Projects
- Questions

Objectives

- Understand the value of Real Time Kafka, Spark and Cassandra Solutions
- Understand the business and technical pieces of how to create integrated solutions
- Discuss how to automate the process

What is Kafka?

- Source Extraction Tool…
- Messaging Tool
- Create Topics that are queries which pull from a source system,
  e.g select * from Customer_Dim
  or select * from Trans_Data
- Incremental/Bulk

What is Spark?

- Written in Scala, Java, or Python
- **Spark Streaming**➔Pull in Kafka Topics and writes out into Cassandra Tables
- E.g. Pull the customer_dim data into Cassandra from the Kafka topic

What is Spark?

- **Spark SQL**→Use for ETL
  - Spark SQL→Using SQL for data transformations, e.g. Group by Sum, Average, function list, etc, incremental / overwrite considerations
  - Spark SQL→ Use Data Frame transformations, e.g. if value = X then y,

What is Cassandra?

- Data Storage→NoSQL→ CQLSH
- NoSQL=SQL with limitations→Different data modeling
- No joins in noSQL, no group bys, no real calcs

What is Cassandra?

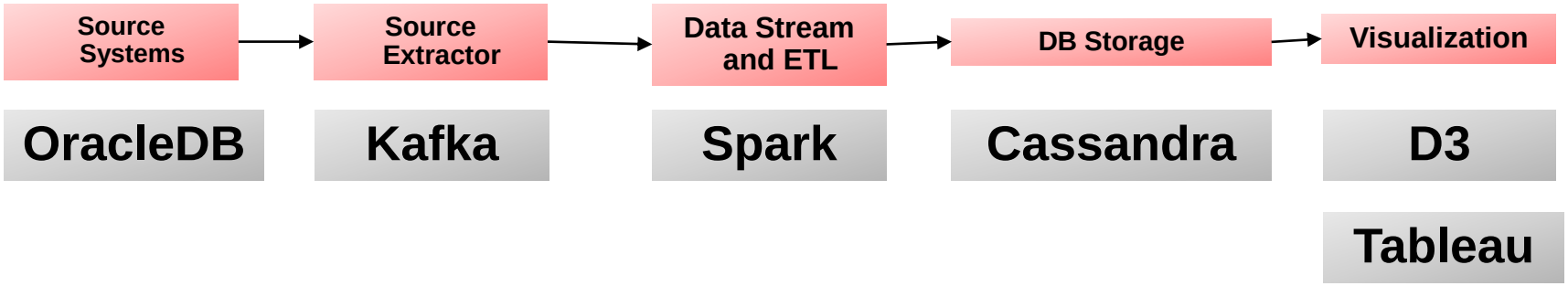
- What do I use this for? Storage and Retrieval
- Highly Available and Fast and Redundant
- Create, Daily, weekly, hourly, etc buckets of tables in Spark SQL
- How is this different than traditional RDBMS?

How to think about processing massive amounts of data

Take massive amounts of data, process it a little at a time….

Similar to drinking a little bit of water at a time to get everything that you need!

Big Data Solution Flow



Kafka, Spark and Cassandra Session Overview

Included:

--Value of Big Data Tools

--How to Integrate Big Data Tools

--Explanation of how Big Data Tools work

Not Included:

--Complex Scala code description

--Review of tons of Kafka/Cassandra settings

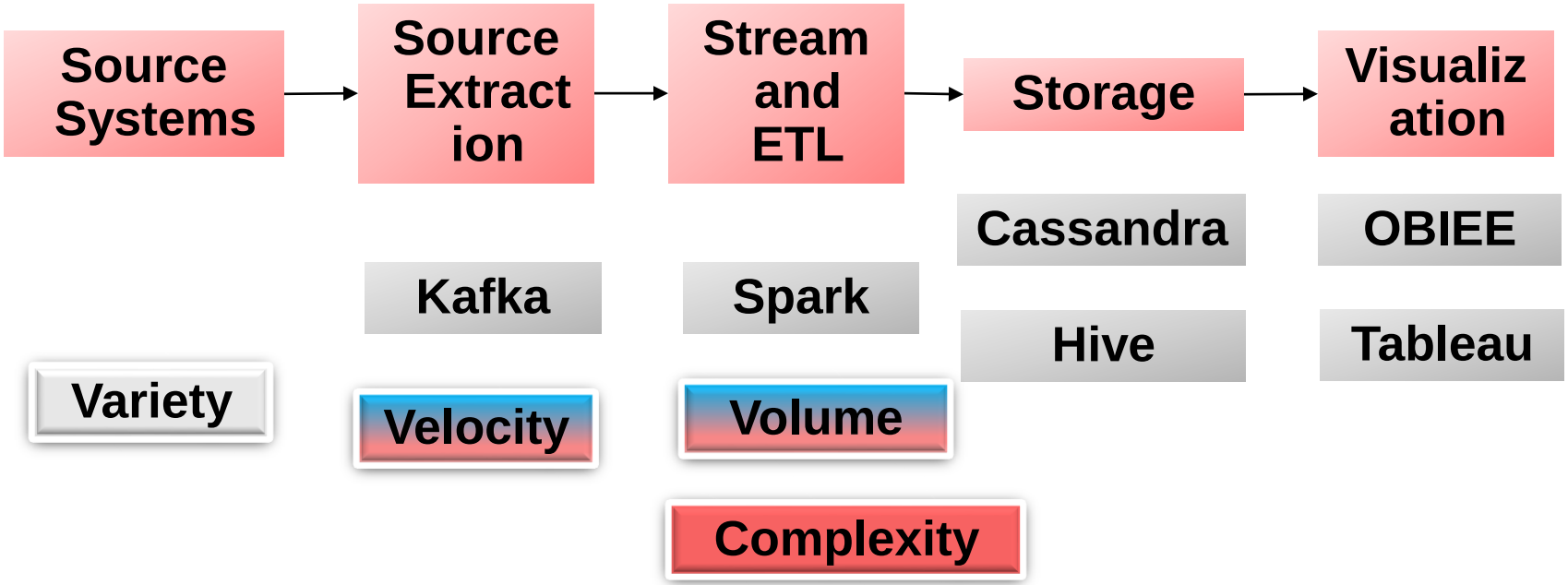
Other Visualization Tools integration with Big Data,

--other Big Data options, Google Big Query

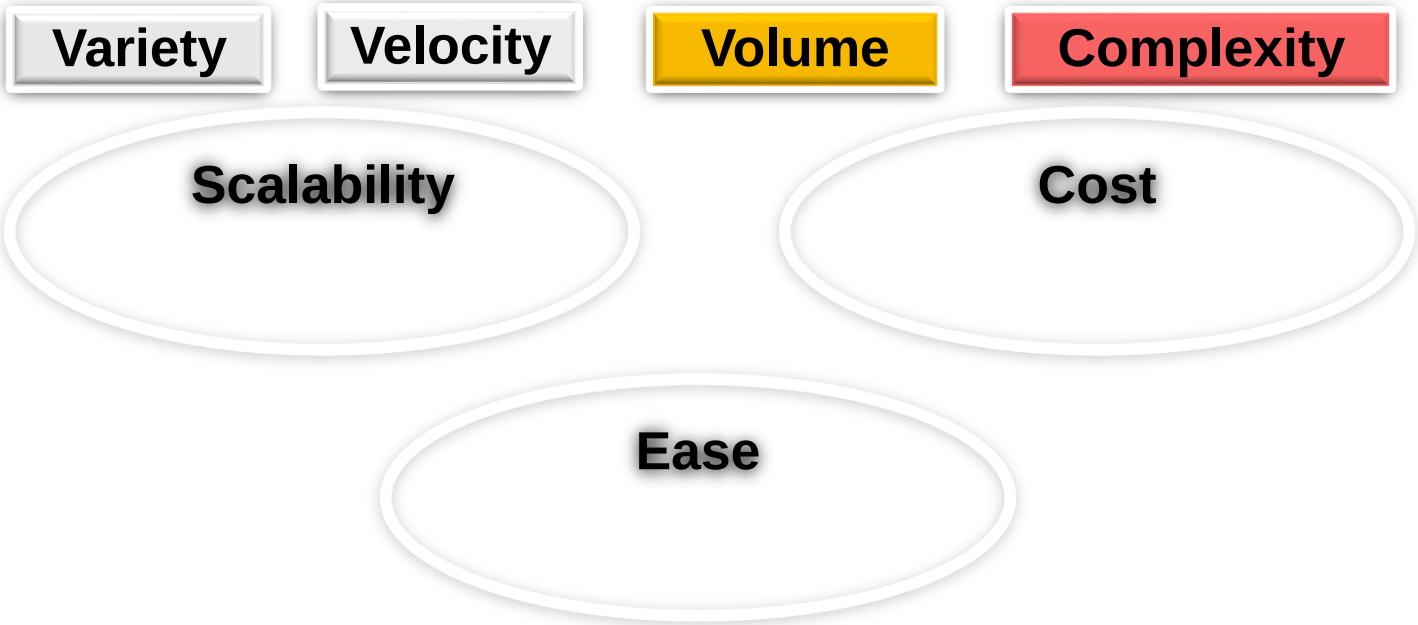
Why do it? The value of Kafka, Cassandra, and Spark

- No licensing fees for solution
- Highly scalable and available solution
- Creating data science environment to analyze data
- Able to apply machine learning algorithms to data for complex analysis
- Real-time, Analytic and Search (solr)
- Pulling in information that wasn’t available—this is a good and a bad thing..

Defining Factors



Defining Factors



Integration Tips: Simplicity

- Keep integration simple and direct.
- Too much information and the purpose of the integration is lost.
- Cut down on information overload.
- Make the end product easy for users!

Kafka Tips

- Try to avoid where clauses—integration modes include incremental and bulk
- Get Data source notices when system goes down will break your topic
- Create simple stream—make sure you try to be incremental and not doing full pulls out of large data sources ☺
- Don’t Crash Source Systems—make sure you understand the system you are querying from
- Bulk can be done incrementally with date time manipulation

Spark Streaming Tips

- Install Cassandra driver and latest version of Spark
- Create simple stream
- Test end to end all of your integration
- Create simple way to split changes.
  (For splitting data we recommend “,,”)

Spark ETL Tips

- Use Spark SQL and DataFrames transformations
- Spark SQL has a nice list of functions to use
  - not all SQL functions are available
  - easy to join tables, perform simple calculations
- DataFrames have some additional pieces
- Set timezone in all servers to UTC or use the timezone in the source systems to be consistent

Spark SQL Functions

- https://spark.apache.org/docs/2.0.2/api/java/org/apache/spark/sql/functions.html
- Typical math functions, concat, string, time

Spark ETL Tips

- SBT—what is it and why to use it?
- Use SBT to create batch ETL jobs, you can parallelize many jobs.
- Use crontab/mesos/yarn to automate the batch process

Cassandra, Spark Integration

- Cassandra Cqlsh is limited, much more so than Spark SQL
- Use Spark SQL to get the data in a
- manageable format for visualization tools

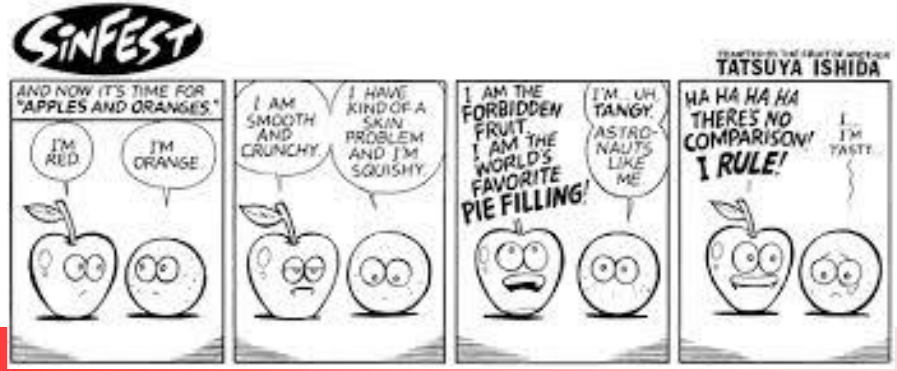
Spark SQL=JDBC Driver to connect to
Cassandra that lets the end user use
Any JDBC tool to analyze Cassandra Data

Cassandra

- CQLSH <> SQL
- CQLSH is fast, replicated, and highly available
- CQLSH can’t join tables
- Is modeled per hour, day, etc, per large query
- Item that you are pulling it for
- Can be accessed using Spark SQL to do all of
- The things you want it to do but can’t.

Kafka, Spark, Cassandra

- Where this can go wrong
  - Different underlying Data sources that don’t integrate
  - Complex ETL that takes too long
  - Data that does not align
  - Project that does not have clear goals of that they want
  - Make sure company can collaborate and work together..



Integration steps of Kafka, Spark, Cassandra

- Integration occurs in Spark.
- Spark has to talk to Kafka—via Spark Streaming
- And also to Cassandra via Spark Streaming
- Luckily connectors exist—we add these to our scala code and to configuration
- Then we can talk to Kafka Topics and write out the results into Cassandra in the same streaming process

Watch out for bugs—it's open source!

Spark connector

Other Big Data Solutions to consider

Kylin—Analytical Tool no Cassandra interface yet

FiloDB—Analytical bought by Apple, for fast olap processing

Solr—for text search/highlighting integrates well with Cassandra

How can we get to OLAP Tree /

Parent-Child Dimensional Analytics?

Key Takeaways

- Understand the business value and how to integrate Kafka, Spark and Cassandra
- Understand the capabilities and how to do projects using Kafka, Spark and Cassandra

BIWA SUMMIT 2018
WITH SPATIAL SUMMIT

THE Big Data + Analytics + Spatial + Cloud + IoT + Everything Cool User Conference

January 30 - February 1, 2018

www.biwasummit.org

Join us around 5:30 today at Oracle Booth if more questions

Please Complete Your Session Evaluation

Evaluate this session in your COLLABORATE app. Pull up this session and tap "Session Evaluation" to complete the survey.

Session ID: 178

COLLABORATE17

TECHNOLOGY AND APPLICATIONS FORUM
FOR THE ORACLE COMMUNITY

Thank you – Please complete evaluation

- Session 178 - Big Data and Oracle Tools Integration
- Kafka, Cassandra, and Spark
- Dan Vlamis
  - dvlamis@vlamis.com
  - 816-781-2880
  - www.vlamis.com
  - @dvlamis
- Jeff Shauer
  - jeffs@jsbusinessintelligence.com
  - 240-205-3042
  - http://www.jsbusinessintelligence.com