Perfecting Your Streaming Skills with Spark and Real World IoT Data

Bob Wakefield
Principal
bob@MassStreet.net
Twitter:
@BobLovesData

Bob's Background

- IT professional 16 years
- Currently working as a Data Engineer
- Education
  - BS Business Admin (MIS) from KState
  - MBA (finance concentration) from KU
  - Coursework in Mathematics at Washburn
  - Graduate certificate Data Science from Rockhurst
- Addicted to everything data

Follow Me!

- Personal Twitter: @BobLovesData
- Company Twitter: @MassStreet
- Blog: DataDrivenPerspectives.com
- Website: www.MassStreet.net
- Facebook: @MassStreetAnalyticsLLC

KC Learn Big Data Objectives

- Educate people about what you can do with all the new technology surrounding data.
- Grow the big data career field.
- Teach skills not products

ACM Kansas City

We’re looking for a speaker willing to talk in deep detail about data engineering challenges their organization is experiencing.

This Evening's Learning Objectives

Learn how to practice your IoT skills with real world IoT data.

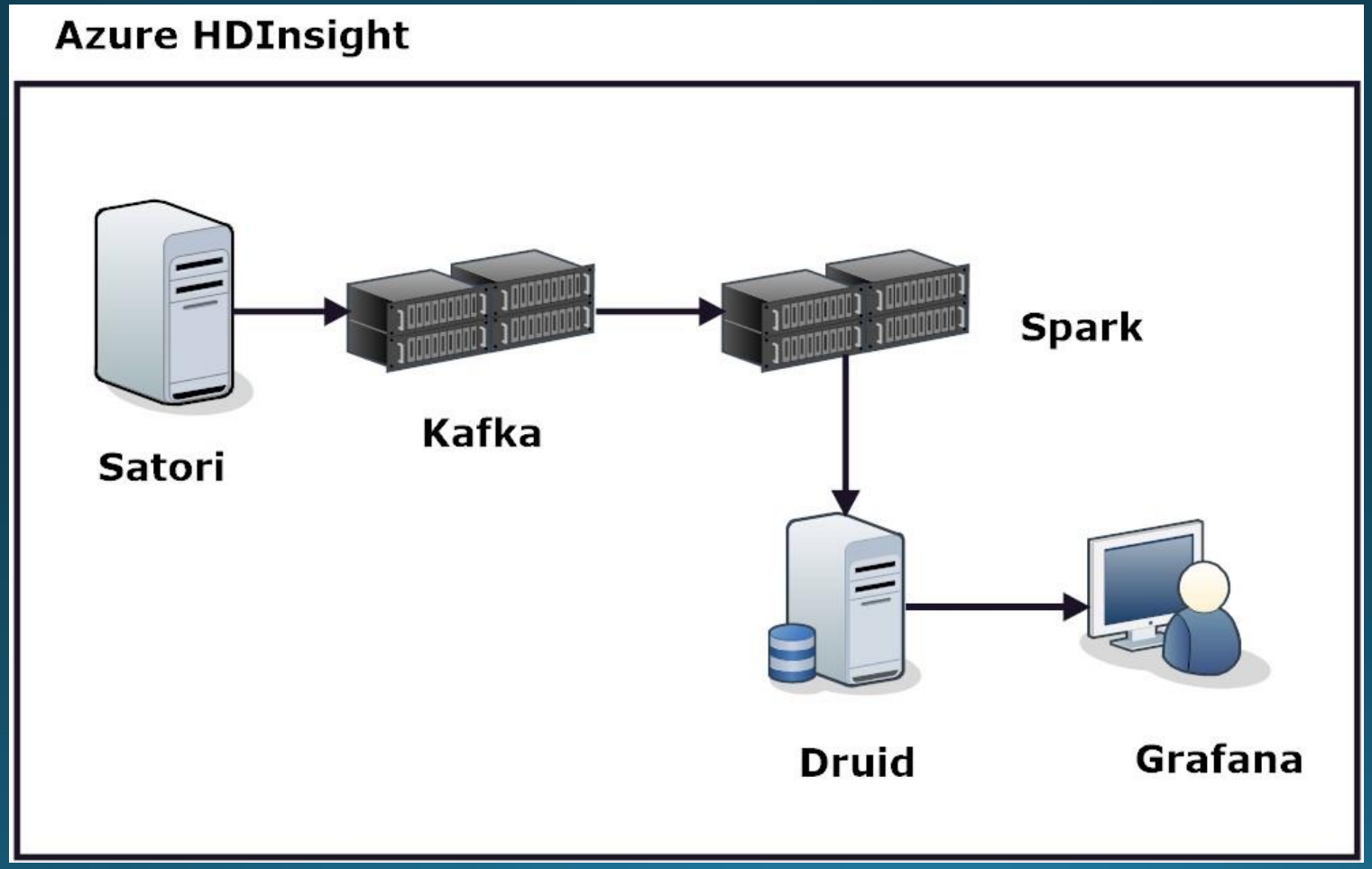
Motivations For This Evenings Discussion

- Getting ready for the opportunities that IoT presents.
- Tired of working with Sandboxes
- Tired of playing with human generated data

Disclaimer

Tonight's presentation is based on a personal hack-a-thon.

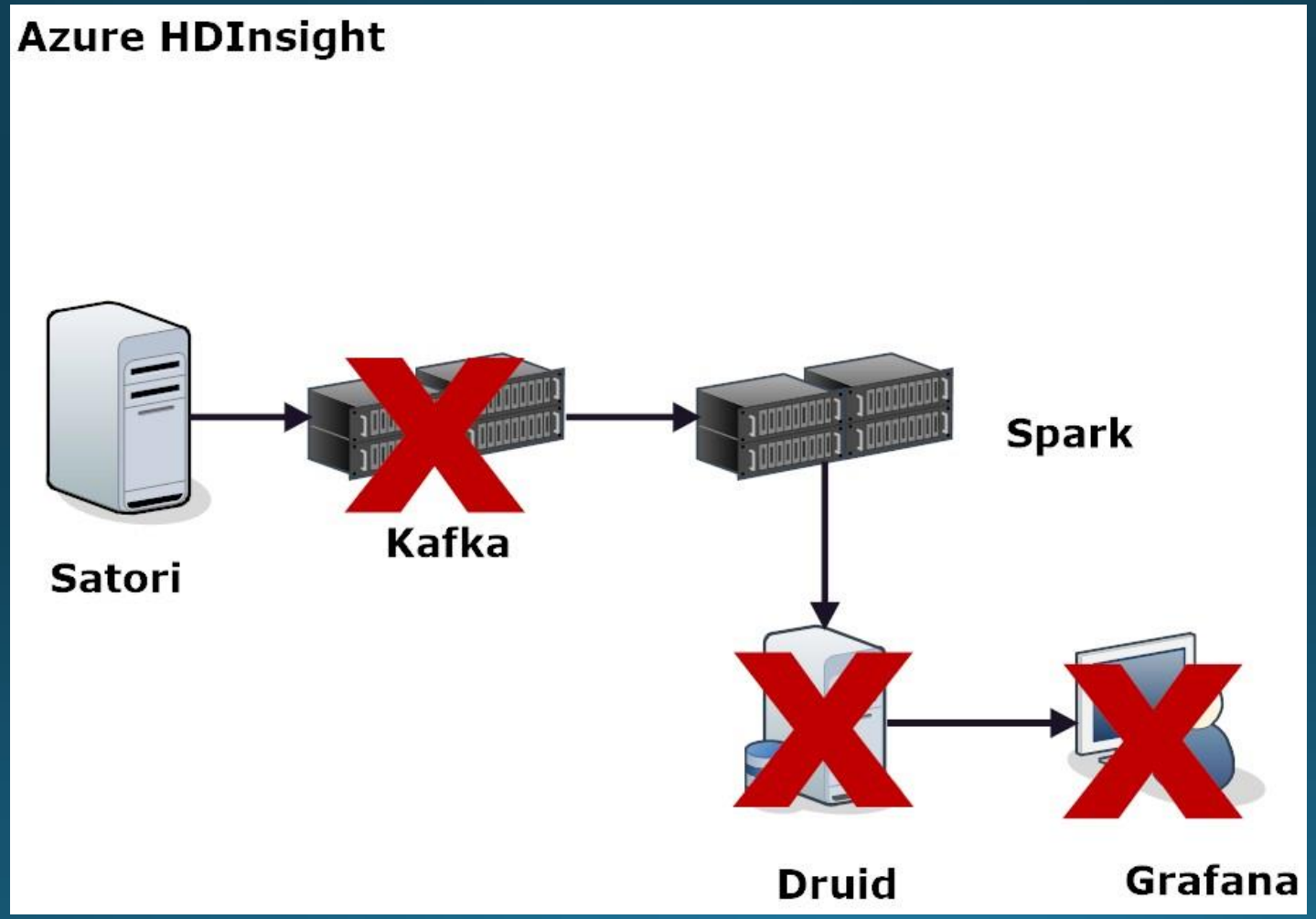
The Original Plan



The Original Plan

- Azure HDInsight
  - Push button cluster
- Kafka
  - Distributed pub/sub messaging system
- Spark
  - Stream Processing framework
- Druid
  - Opensource OLAP NoSQL real time database
- Grafana
  - IoT Dashboard

The New Plan



All Material Can Be Downloaded from GitHub

MassStreetAnalytics/iot-with-satori

Azure HDInsight

- Button push Hadoop cluster
- Cloud version of Hortonworks Data Platform
- You can spin up different types of clusters
  - Plain Hadoop
  - Spark
  - Hbase
  - R Server
  - Storm
  - Real Time Hive
  - Kafka

Azure HDInsight

- Each cluster type comes with the following
  - Ambari
  - Avro
  - Hive and Hcat
  - Mahout
  - MapReduce
  - Oozie
  - Phoenix (?) (I don’t normally play with Hbase)
  - Pig
  - Sqoop
  - Tez
  - Yarn
  - ZooKeeper

Azure HDInsight

Let's stand up a cluster!

Azure HDInsight

Make sure you blow away the resource group!

The Case for Learning IoT

- IoT represents a significant business opportunity
  - Still on the wrong side of the hype cycle
  - Not sure why
  - Might be due to hardware requirements
  - Just a matter of time
- IoT opens up an new world of applications
- IoT use cases
  - Has the potential for ubiquitousness

An IoT Case Study
Life Alert

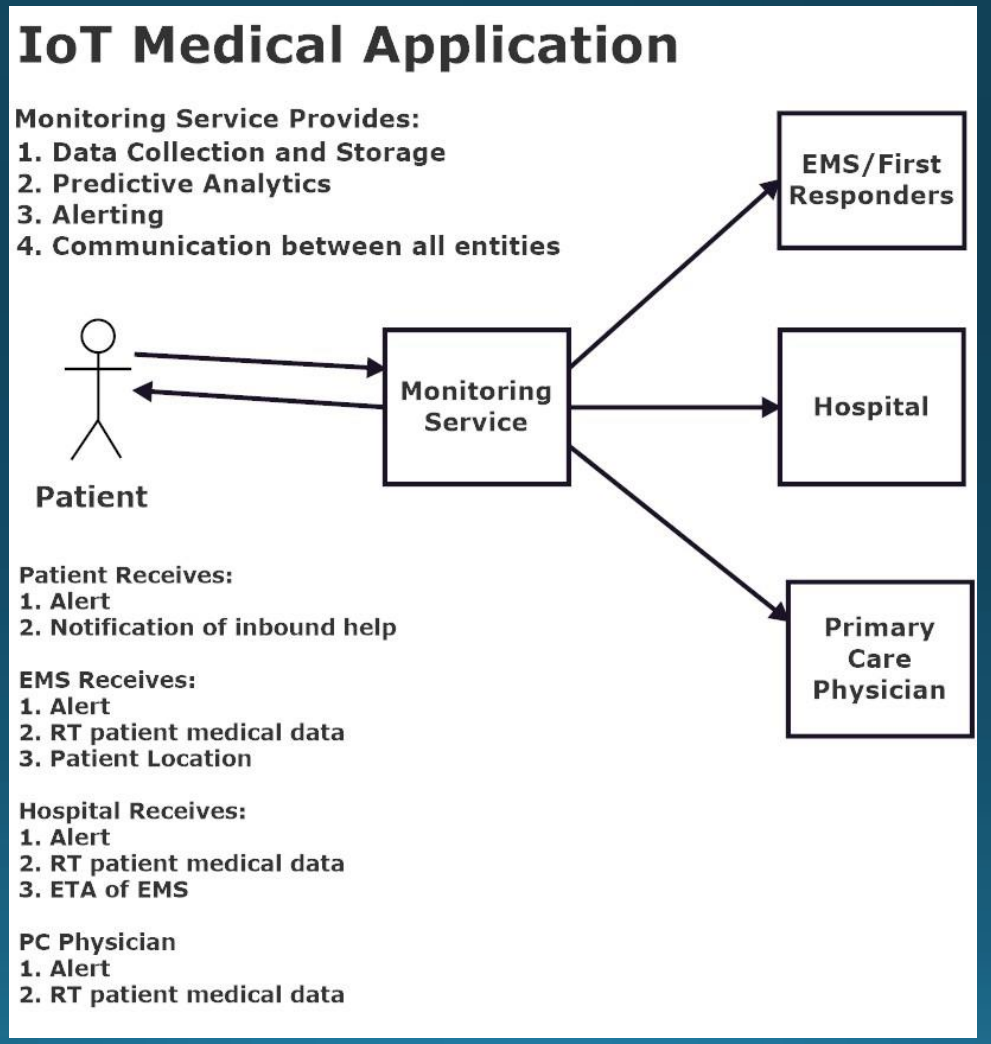
Fatal flaw (no pun intended) .

Patient has to be conscious to activate.

An IoT Case Study

Software is easy.

Challenge: Creating a device with a form factor that provides function but doesn't get in the way of the patients life.



Satori

- www.satori.com
- Open Streaming Data Platform
  - Appears to be in tech preview
  - So far appears to be free on the sub side
- Allows people to pub/sub to data streams
- A lot of municipal and science streams
- Not just IoT data

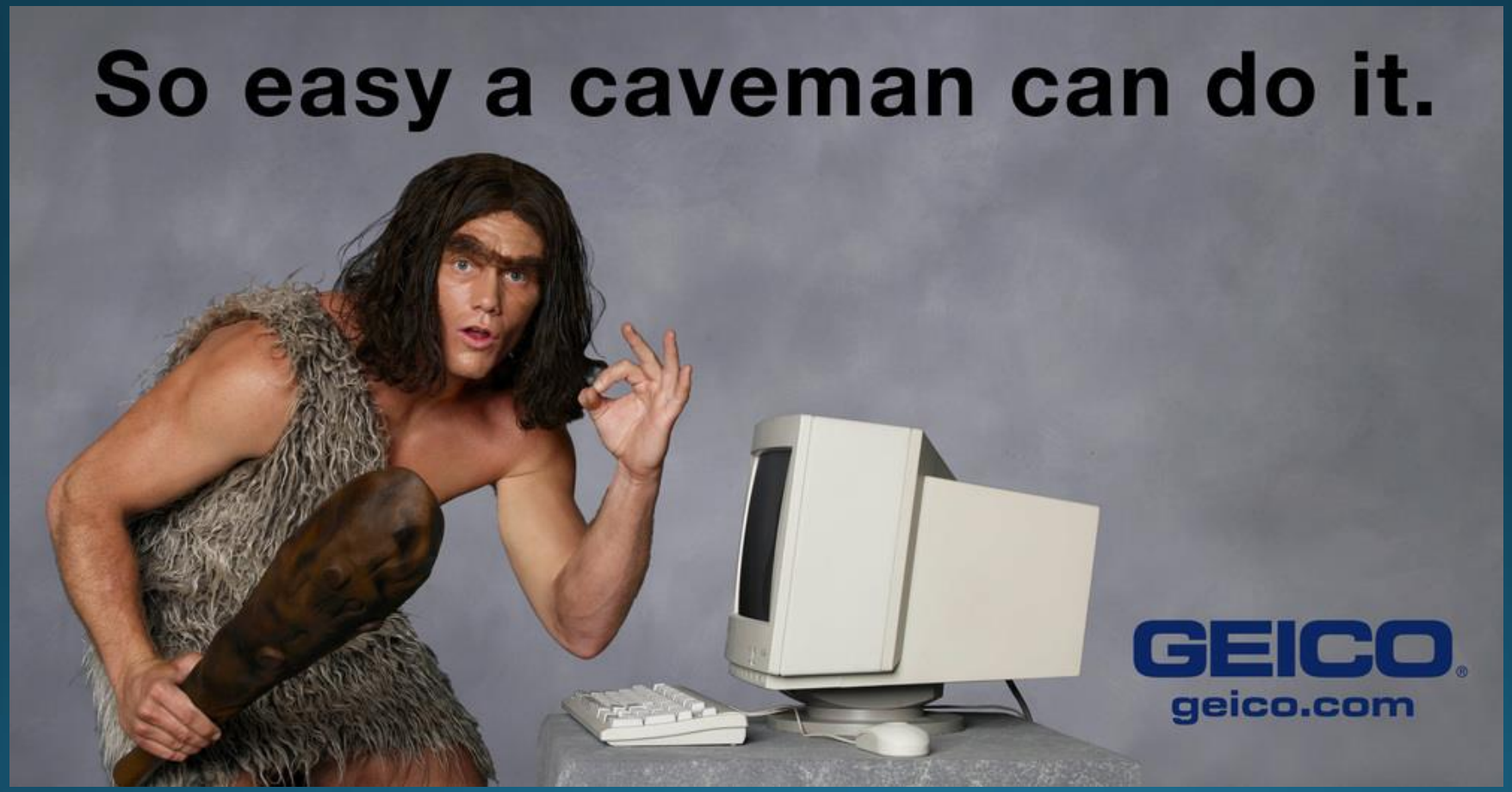
Satori

- You need an SDK to build stuff with Satori
- You can use your favorite build tool to import the necessary classes

Satori

Let’s look at some code!

Spark Structured Streaming….

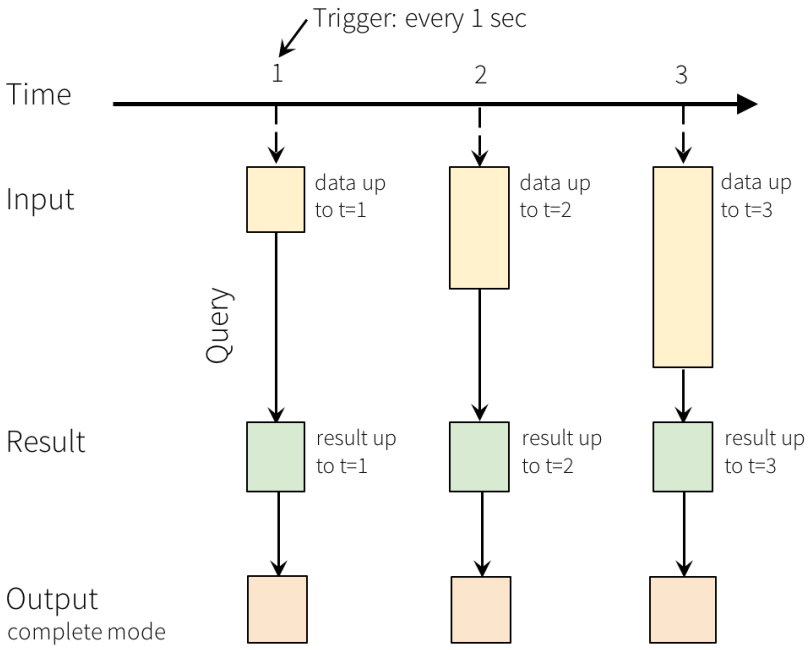


```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.SparkSession

val spark = SparkSession.builder.appName("StructuredNetworkWordCount").getOrCreate()

import spark.implicits._

// Create DataFrame representing the stream of input lines from connection to localhost:9999
val lines = spark.readStream.format("socket").option("host", "localhost").option("port", 9999).load()

// Split the lines into words
val words = lines.as[String].flatMap(_.split(" "))

// Generate running word count
val wordCounts = words.groupBy("value").count()

// Start running the query that prints the running counts to the console
val query = wordCounts.writeStream.outputMode("complete").format("console").start()

query.awaitTermination()
```

Spark Structured Streaming

- Represents a VAST improvement over previous stream processing frameworks
  - Storm
  - Flume
  - Flink
  - Samza
  - Spark Streaming
- Allows you to create stream processes without having to think much about it.

Spark Structured Streaming



Programming Model for Structured Streaming

Spark Structured Streaming

- If you can make batch jobs with Spark SQL, you can make Structured Streaming Jobs.
- So new there is little to no literature on the topic.
  - Structured Streaming Programming Guide is your best bet.
- Exactly once delivery semantics out of the box.
- Limited number of built in sources
  - Socket
  - Kafka
  - File

Spark Structured Streaming

- Fairly unlimited on sinks
  - JSON
  - ORC
  - Parquet
  - CSV
  - Database table

Spark Structured Streaming

- Things you can do with Structured Streaming
  - SQL Operations
    - Grouping/aggregations
    - Filtering
    - Joins (one DF has to be static)
    - Things you can't do make no sense in a streaming context
      - Limit
      - First N rows
- Operations over sliding windows
- Handling late data with watermarking

Spark Structured Streaming

Let’s look at some code!

Code Challenge

- Convert the socket server into a Kafka Producer
- Find a streaming source that's simpler in structure than the weather data and use it
- Attempt to parse the weather data and load it into a case class.
- Try and build the rest of the application