INTRODUCTION TO

APACHE:

BIG_DATA

NORTH_AMERICA
AGENDA

PROBLEM: PROLIFERATION
ACTIVITY STREAMS
APACHE STREAMS
COMPATIBILITY
SCHEMAS
RESOURCES
PROBLEM: PROLIFERATION!

SILOS
STANDARDS
SCHEMAS
SDKS
DATABASES
FRAMEWORKS
RUNTIMES
CHALLENGING TO GET A COMPOSITE PICTURE OF A PERSON OR ORGANIZATION. DATA RESIDES IN MANY SYSTEMS THAT ARE NOT EASILY INTEGRATED.
STANDARDS

NO UNIVERSALLY ADOPTED STANDARD FOR STRUCTURING SOCIAL PROFILES

NO STANDARD FOR TRANSMITTING ACTIVITIES ACROSS DATA SILOS

THIS IS TRUE ACROSS WEB SITES, AS WELL ACROSS ENTERPRISE APPLICATIONS
MOST DATA SILOS
MAKE MINIMAL (IF ANY) EFFORT TO PROMOTE INTEROPERABILITY
DON’T PUBLISH MACHINE-READABLE API SPECS
DON’T PUBLISH MACHINE-READABLE SCHEMAS
SDKS

MANY DATA SILOS RECOMMEND USAGE OF ONE OF THEIR SDKS TO USE THEIR DATA SERVICES

HOWEVER THESE SDKS IMPOSE THEIR PREFERRED LIBRARIES WITHOUT REALLY MAKING DEVELOPMENT EASIER
DATABASES

UNPRECEDENTED RANGE OF CHOICES
DEVELOPERS HAVE PREFERENCES
ORGANIZATIONS HAVE RESTRICTIONS

MANY APPLICATIONS REQUIRE A POLYGLOT ARCHITECTURE TO SCALE.
WHO WANTS TO LEARN THE PROTOCOLS AND ASSUMPTIONS OF NEW DATABASES CONSTANTLY?
FRAMEWORKS

CAN BE VERY HELPFUL WHEN BUILDING SCALABLE SYSTEMS

BUT:
THEY ENFORCE CONVENTIONS AND HAVE ANNOYING CONSTRAINTS.
THEY LEAD TO LOCK-IN, UNLESS YOUR TEAM IS EXTRA-ORDINARILY VIGILANT.
RUNTIMES

THE CLOUD MAY BE CHEAPER BUT RUNTIME-SPECIFIC VARIATION IMPACTS THE WAY WE:

PACKAGE
DEPLOY
CONFIGURE
MONITOR

RUNTIMES LEAD TO LOCK-IN, UNLESS YOUR TEAM IS EXTRA-ORDINARILY VIGILANT.
ACTIVITY STREAMS OBJECTIVES

LANGUAGE AGNOSTIC
CROSS-APPLICATION INTEROPERABILITY
SUPPORT FOR MULTIPLE SCHEMAS
STREAM FEDERATION
STREAM FILTERING
ACTIVITY STREAMS BASICS

NORMALIZED FORM FOR ENTITIES AND EVENT
<ACTOR> DID <VERB> WITH <OBJECT> (TO <TARGET>) AT <PUBLISHED>

OBJECT TYPES
PERSON, ORGANIZATION, IMAGE, VIDEO, ETC...

VERBS
POST, SHARE, LIKE, ETC...
IMPLEMENTATION CHALLENGES

ADOPTION
- Industry support has been tepid at best

AMBIGUITY
- The spec itself is open to interpretation

EXTENSIONS
- The spec rightly allows for arbitrary extensions

FLEXIBILITY
- As a result, activities from any two providers are just barely interoperable

VALIDATION
- Data correctness or coherence is not covered by spec
APACHE STREAMS

A LIGHTWEIGHT (YET SCALABLE) FRAMEWORK FOR ACTIVITY STREAMS
AN SDK FOR BUILDING DATA-CENTRIC JVM APPLICATIONS
A SET OF PATTERNS FOR BUILDING RELIABLE, ADAPTABLE, DATA PROCESSING PIPELINES
PHILOSOPHIES

BE DATABASE AGNOSTIC
BE RUNTIME AGNOSTIC
ENFORCE TASK AND DOCUMENT SERIALIZABILITY
DOCUMENTS AS THE CORE UNIT OF PROCESSING
SUPPORT ANY TYPE OF DOCUMENTS AND ARBITRARY METADATA
ENCOURAGE EXPLICIT SPECIFICATION OF DOCUMENTS VIA JSON SCHEMA AND XML SCHEMA
ASSIST WITH CONVERSION TO AND FROM ACTIVITYSTREAMS
INTERFACES

PROVIDER
- TASK THAT SOURCES DOCUMENTS FOR THE STREAM, LIKELY IN THEIR ORIGINAL DATA FORMAT.

PROCESSOR
- TASK THAT TRANSFORMS DOCUMENTS, PERHAPS WITH A SYNCHRONOUS CALL TO AN EXTERNAL SYSTEM.

PERSIST READER
- TASK THAT SOURCES DOCUMENTS FROM A FILE SYSTEM OR DATABASE.

PERSIST WRITER
- TASK THAT SAVES DOCUMENTS TO A FILE SYSTEM OR DATABASE.
COMPATIBILITY DIMENSIONS

PROVIDERS
PERSISTANCE
PIPELINES
RUNTIMES
SCHEMAS
COMPATIBILITY:
PERSISTANCE

BUFFER (FILE SYSTEM)
CASSANDRA
ELASTICSEARCH
GRAPH (NEO4J)
HBASE
HDFS
MONGODB
KAFKA
KINESIS
S3
COMPATIBILITY:
RUNTIME FRAMEWORKS

DOCKER
DROPWIZARD
PIG
SPARK
STORM
COMPATIBILITY:
RUNTIME ROADMAP

CRUNCH
FLINK
LOGSTASH
NIFI
SAMZA
SPARK STREAMING
TWILL
COMPATIBILITY: SCHEMAS

SCHEMATA ARE:
THE PRESENCE AND ABSENCE OF FIELDS AND STRUCTURE DIFFERENT FROM CLASS AND FROM FORMAT

STRATEGIES FOR SCHEMA MANAGEMENT
MANY-TO-MANY
MANY-TO-MINE
MANY-TO-ONE

SCHEMA-RELATED CHALLENGES
SCHEMA MANAGEMENT: MANY-TO-MANY

FOR EVERY PROVIDER AND TYPE: MAP AND CONVERT TO COMPATIBLE TYPES FROM ALL OTHER PROVIDERS

THIS IS THE DEFAULT MODALITY FOR DATA AND IT SUCKS
SCHEMA MANAGEMENT: MANY-TO-MINE

SPECIFY INTERNAL TYPES, THEN FOR EVERY PROVIDER AND TYPE:

ASSESS, ALIGN AND CONVERT TO PREFERRED INTERNAL REPRESENTATION

THIS IS BETTER, BUT IT FAILS AS SOON AS WE WANT TO INTEROPERATE WITH OTHER
DEPARTMENTS OR ORGANIZATIONS WHO ARE ALL USING THEIR OWN INTERNAL SCHEMAS

EXPECT TO CHANGE YOUR INTERNAL SPEC RELATIVELY OFTEN IN EARLY STAGES,
MEANING YOU PROBABLY HAVE TO EITHER

UPGRADE YOUR DATA OR

GUARANTEE BACKWARD COMPATIBILITY IN-APPLICATION
SCHEMA MANAGEMENT:
MANY-TO-ONE

FOR EVERY PROVIDER AND TYPE:
A COMMUNITY DEDICATED TO THE INTER-
OPERABILITY OF THAT DATASET Sorts OUT A
REASONABLE MAPPING TO A RELATIVELY STATIC
PUBLIC SPECIFICATION
WHERE THE EXISTING PUBLIC SPECS ARE
INADEQUATE:
THE COMMUNITY CAN FIND A WAY TO ESTABLISH
COMPATIBILITY VIA CONVENTION. OPEN-SOURCE
COMMUNITIES AND STANDARDS BODIES CAN
COLLABORATE FOR BENEFIT OF ALL
SCHEMA CHALLENGES:
SHARING

BUSINESS-AS-USUAL:
SCHEMAS ARE OFTEN IMPLICIT, SHARED VIA UNSTRUCTURED WEB DOCUMENTATION AND LANGUAGE SPECIFIC SDKS

STREAMS:
STREAMS SOURCE CODE CONTAINS JSON AND XML SCHEMAS FOR MANY SUPPORTED PROVIDERS ANYONE CAN IMPORT OR EXTEND THESE SCHEMAS (VIA HTTP!)
SCHEMA CHALLENGES: DATE-TIMES

BUSINESS-AS-USUAL:
HERE’S A STRING, HAVE FUN!

STREAMS:
EVERY LIBRARY ON THE CLASSPATH DECLARES ITS PREFERRED FORMAT(S)
FRAMEWORK RESOLVES ANY KNOWN FORMAT AND USES JODA TO CONVERT TO RFC3339
SCHEMA CHALLENGES:
VERSIONING

BUSINESS-AS-USUAL:
SCHEMAS CHANGE AS PRODUCT AND API FEATURES EVOLVE, AND EVERYONE JUST MUDDLES THROUGH.

STREAMS:
SCHEMAS GET PUBLISHED WITH EVERY RELEASE AND EVERY SNAPSHOT FOR BENEFIT OF THOSE RESPONSIBLE FOR DEPENDENT LIBRARIES CHANGES GET DESCRIBED IN RELEASE NOTES UPDATES TO UNIT AND INTEGRATION TESTS

APACHE:
BIG_DATA
NORTH_AMERICA
SCHEMA CHALLENGES:
IDE SUPPORT

BUSINESS-AS-USUAL:
IMPORT OUR SDK OR GTFO

STREAMS:
ALL STREAMS TYPES HAVE A SERIALIZABLE POJO REPRESENTATION
IMPORTABLE WITH MAVEN TO SPECIFIC VERSION
CONVERTIBLE TO ANCESTOR, SIBLING, AND CHILD TYPES WITH A CAST
CONVERTIBLE TO OTHER TYPES WITH A ONE-LINER
SCHEMA CHALLENGES:
IMPORTS

BUSINESS-AS-USUAL:
EVERY SERVICE IS AN ISLAND

STREAMS:
‘EXTENDS’ CAPABILITY OF JSON SCHEMA ALLOWS FOR EMERGENCE OF A WEB OF RELATED TYPES
DESCRIBE YOUR OBJECTS AS A DELTA TO BASE SCHEMAS OR A MASHUP OF SEVERAL UNDECLARED FIELDS PROPAGATE BY DEFAULT
SCHEMA CHALLENGES:  
CONVERSION

BUSINESS-AS-USUAL:
EITHER GET TOO MUCH TYPE SAFETY OR NONE, TAKE YOUR PICK
IF YOU'RE LUCKY, FRAMEWORK HELPS WITH SERIALIZATION AND COMPRESSION
STREAMS:
INCLUDES MULTIPLE TYPE CONVERSION OPTIONS, AVAILABLE AS PROCESSORS FOR YOUR
STREAMS OR SINGLETON UTILITY CLASSES TO EMBED IN YOUR CODE
   JACKSON CONVERSION
   HOCON CONVERSION
   VIA JAVA/SCALA
RESOURCES

WEBSITE
   HTTP://STREAMS.INCUBATOR.APACHE.ORG/

SOURCE CODE
   HTTPS://GITHUB.COM/APACHE/INCUBATOR-STREAMS

DOCUMENTATION
   HTTP://STREAMS.INCUBATOR.APACHE.ORG/SITE/0.2-INCUBATING/STREAMS-PROJECT/INDEX.HTML

EXAMPLES
   HTTPS://GITHUB.COM/APACHE/INCUBATOR-STREAMS-EXAMPLES

EXAMPLES DOCUMENTATION
   HTTP://STREAMS.INCUBATOR.APACHE.ORG/SITE/0.2-INCUBATING-SNAPSHOT/STREAMS-EXAMPLES/INDEX.HTML

APACHE:
BIG_DATA
NORTH_AMERICA