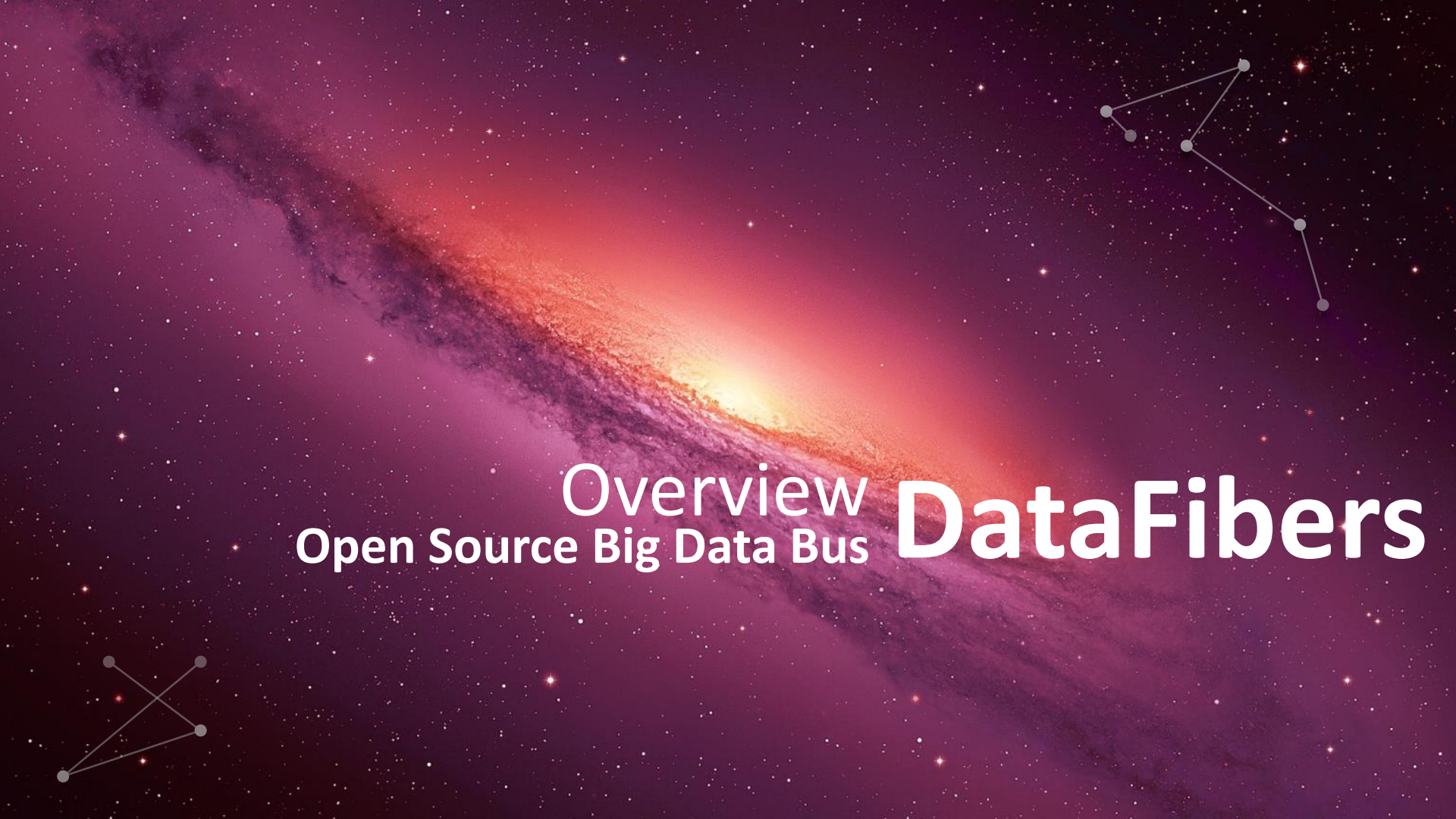

Overview
Open Source Big Data Bus
DataFibers

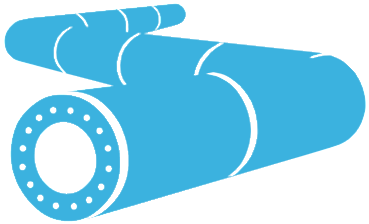
Overview

- **DataFibers** – short as Fibers, a open source implementation of enterprise big data bus.

- DataFibers simplifies the roadmap for enterprise to rock with big data.

History of Bus Architecture

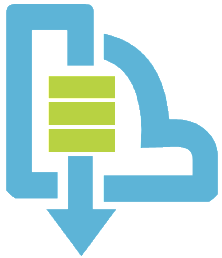
- A pipeline, backbone, highway, trunk, bus, etc.
- A naturally excellent design pattern for information sharing and management



1870's first trunk pipeline

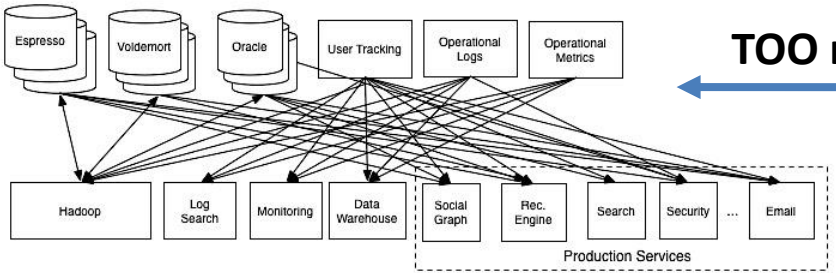
1910's first highway

1990's Internet & WWW



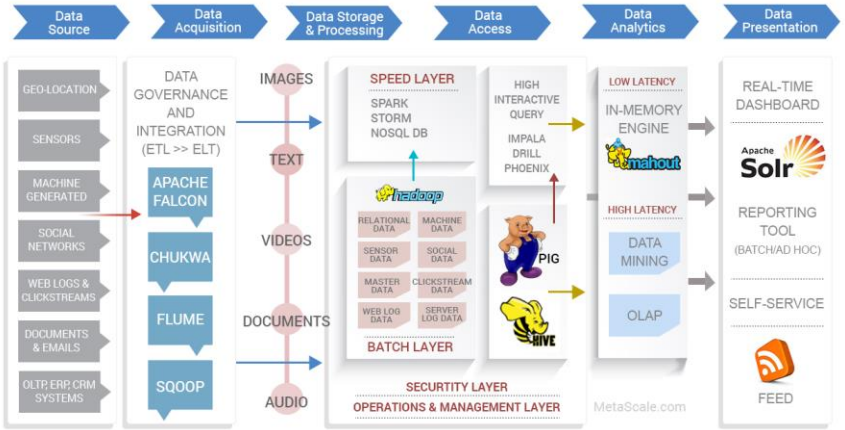
2000's ESB

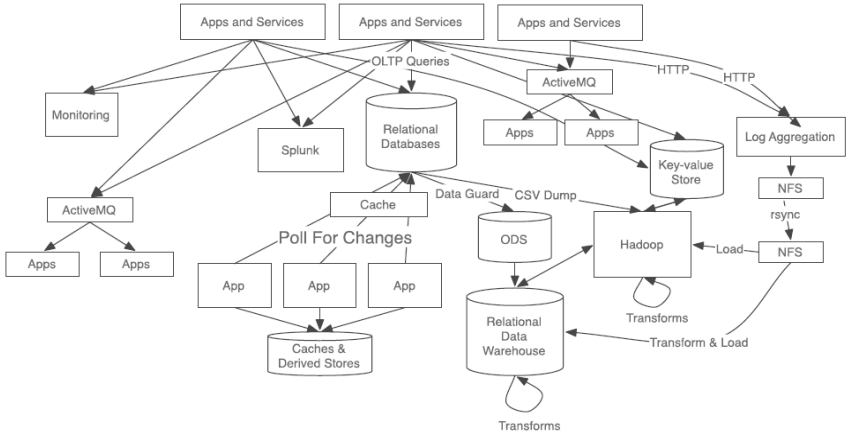
The Pain Before Data Bus



TOO much pipes & work

TOO complex





TOO many steps & layers, and too slow

The Return of the King

Wait a minute, a new king **DataFibers – An Enterprise Big Data Bus**

A Bus

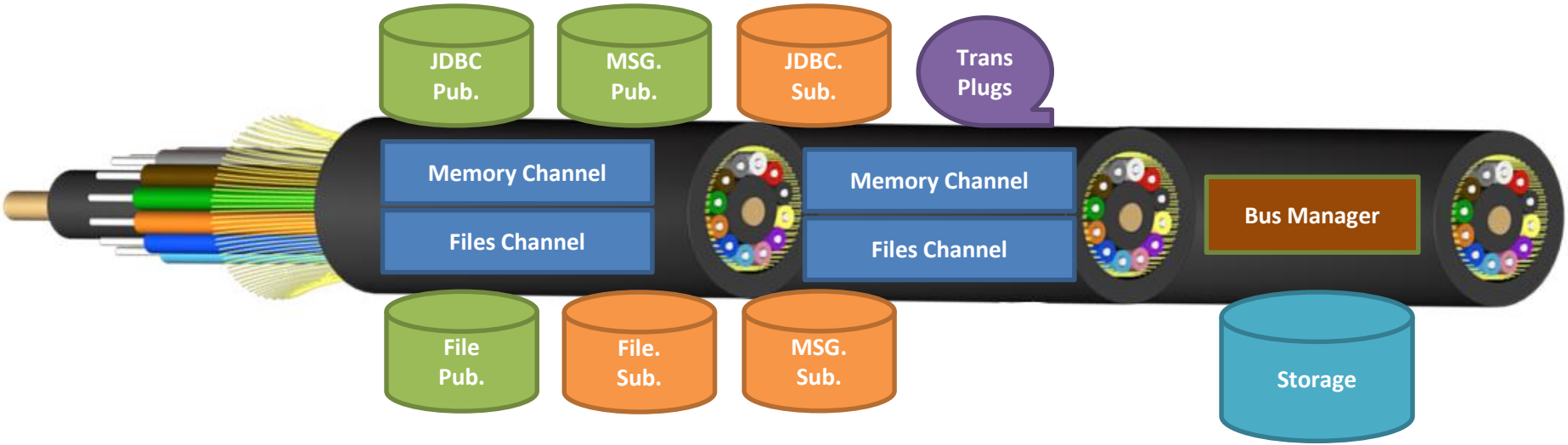
A Data Bus

A Big Data Bus

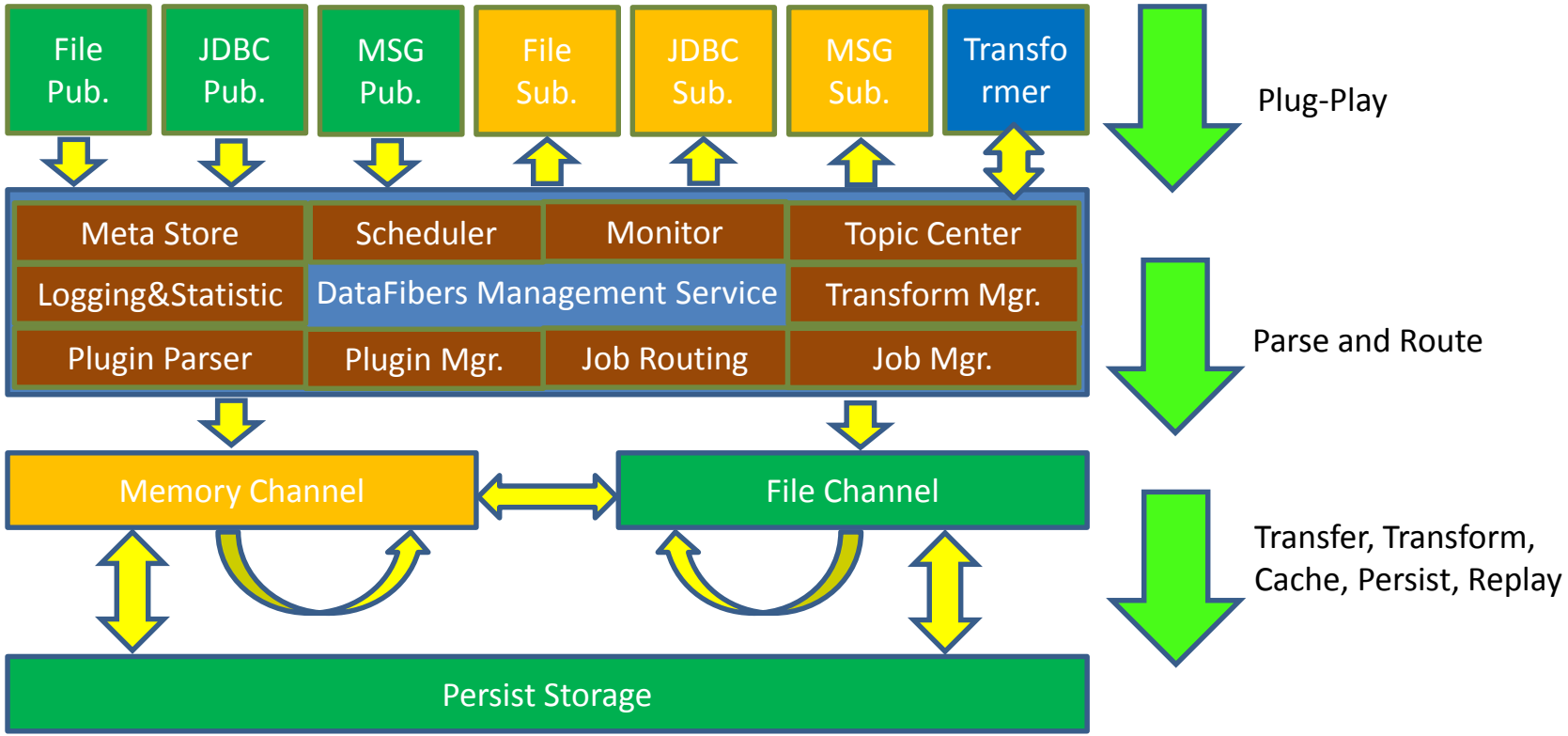
Built for Enterprise Big Data

DataFibers Logic Architecture

- Publisher, Subscriber, Transformer,
- Memory Channel, File Channel
- Meta and Data Storage



DataFibers Physical Architecture



DataFibers Technical Stacks - TBD

- Apache Hadoop, Alluxio (Tachyon), Kudu
- Spark, Flink, Beam
- Hive, MapReduce
- Kafka, Vertx
- Elastic, MongoDB, HBase, PostgreSQL

DataFibers Features

- Dynamic and transparent routing data processing
- Transformation between jobs
- Simple and powerful (especially on transformation) API (vs. Spark?)
- Data subscribe, pull, and push
- Topic exploration, management, and subscription
- Support batch, stream, real-time, and hybrid data processing
- Data cache, replay, reprocess
- Messaging metadata for data discovery and optimized access
- Inter-bus connector, bus-hub and data market place

DataFibers Use Cases

- Message queue and streaming
- Batching data processing
- Hybrid data processing
- Streaming and batching data transformation, such as lookup, join, filter
- Speed up data discovery without pulling or migrating data into single storage
- A logical place overview for all of the enterprise data
- An outbox unified data processing and accessing framework
- A full life cycle of data – collected, cached, transformed, replayed, reused

DataFibers User Story

- An investment bank of North America region has regular data processing in both batch and stream. The core stream data is mainly about how are customers buying and selling stocks from its stock portal.
- External streams, such as Tweeter and Facebook data, are also being used to calculate and recommend the best stocks.
- All of history data of stock is ingested to the Hadoop as data hub and data warehouse ease of stock recommendation in real time.
- A real-time recommendation of stocks is provided to each individual customers based on his current P&L, history transaction, market trends, etc.
- The same type of data is also replicated in real-time to customer care service database to keep track the performance of their agent and recommendation service.
- An updated dashboard is generated by combining history data and current data for senior leadership's decision making in agile mode.
- A data science team should have access to all metadata and data sets in order to discover the valuable patterns and strategy for algorithm trading and stock price prediction.
- Exchanging data among internal bank system, such as MDM, ERP, CRM, Finance, Loyalty, Audit need to be established.
- Two other region business operations which have their own data hub or lake need to exchange data and metadata in efficient way with North America region.

DataFibers User Story …

Dot end lines are either batch or real-time stream

Internal Discovery Portal

BI Dashboard Portal

Data Auditing

Other DataFibers Instance

Key-Value Store

Memory Cache

3rd Party Feeds

transform

Iterative Training

Lookup

Enterprise Big Data Bus - DataFibers

NA Data Hub

NA MDM

NA CRM

NA EDW

Asia Data Hub

EU Data Hub

Our Challenges

- Start from very beginning
- Competition from ESB, Data Hub, ETL in Cloud
- Ecosystem has rapid changes and too many choices
- Real-time/streaming is still new and involving
- Integration and extension
- Enterprise level security
- Use cases, stories, sponsors

Our Opportunities

- Stand on the shoulders
- Big data processing is too complex. A simple and unified pattern is expected
- A logic view of big data is always more practical than physical
- One data processing framework for all type of processing data
- The vision and strategy for the full life cycle of enterprise data
- Decoupling, optimized, sharing, extensible patterns
- A roadmap from data hub, data lake to data ocean and market

Project Management

- PMC – Active steer the project direction
- Advisors – Seasonal consultant
- Committers – active monthly
- Contributes – ad-hoc commit
- Release Manager – ad-hoc by release timeline

We follow the agile scrum model.
At this time, PMC will play the role of PO and RM plays Scrum Master

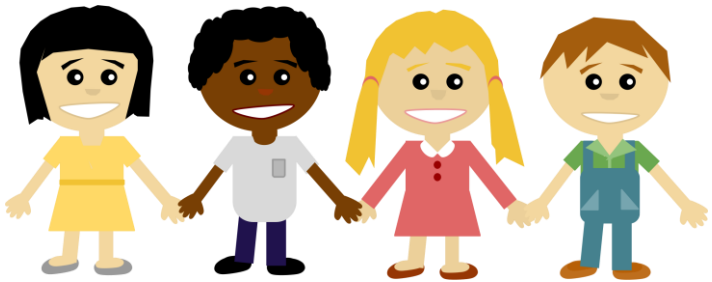
Core Roles and Skills Needed

- Java | Scala full stack developer
- Spark | Flink | Kafka engineer
- Integration | Infrastructure engineer
- Front End | Dashboard Designer
- Seasonal technical writer | Blogger
- User story sponsors | business development
- Program manager | Advisors



What We Offers

- An opportunity to make things big and different
- Experience to touch deep to the big data ecosystem
- An chance to work with top-notch big data professionals
- Open source communities
- Start-up opportunities
- Advices, help, reference, friendship, etc.



How to Participant

- Get to know about the project at datafibers.org | datafibers.com
- Watch and Star us in GitHub
- Fork and Pull Request when you have ideas to contribute
- Contact Us for participant at datafibers@gmail.com
- Join our discussion and ask questions at datafibers@googlegroups.com
- Hear our news and events @

  data.fibers datafibers1

- Our Meetup is incoming

www.datafibers.org
datafibers@gmail.com
datafibers@googlegroups.com

Thank You

QUESTIONS