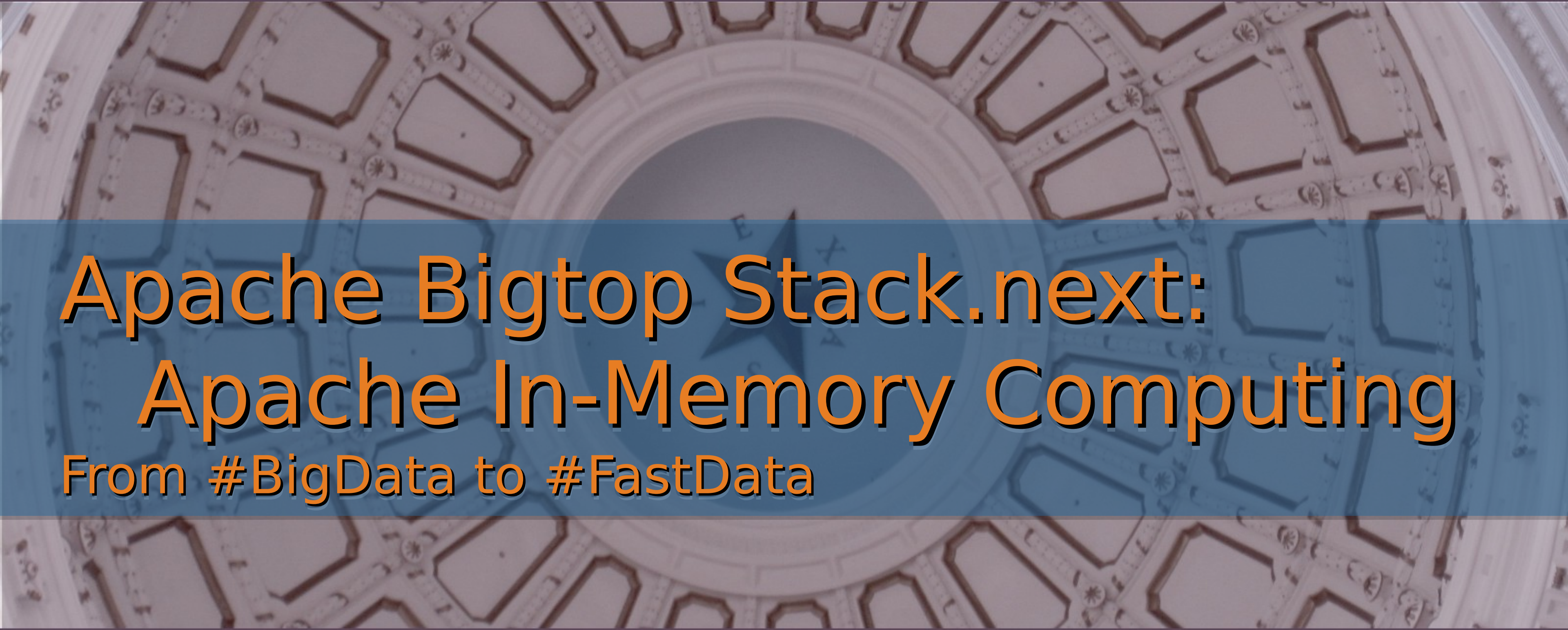
Dr. Konstantin Boudnik
Apache Bigtop PMC Chair

cos@apache.org

APACHECON
NORTH AMERICA



Apache Bigtop Stack.next: Apache In-Memory Computing

From #BigData to #FastData

Solving the complexity

#BigData

Apache Bigtop primer

- A project, environment, and a phylosophy to:
  - Define and create software stacks (think Debian)
  - Deploy and validate actual software in the real world
  - Configuration management
- Guarantees of consistency and compatiblity
- Empirical vs Rational
  - don't rely on someone's hearsay
  - don't assume an environment: contol it

One stack to rule them all

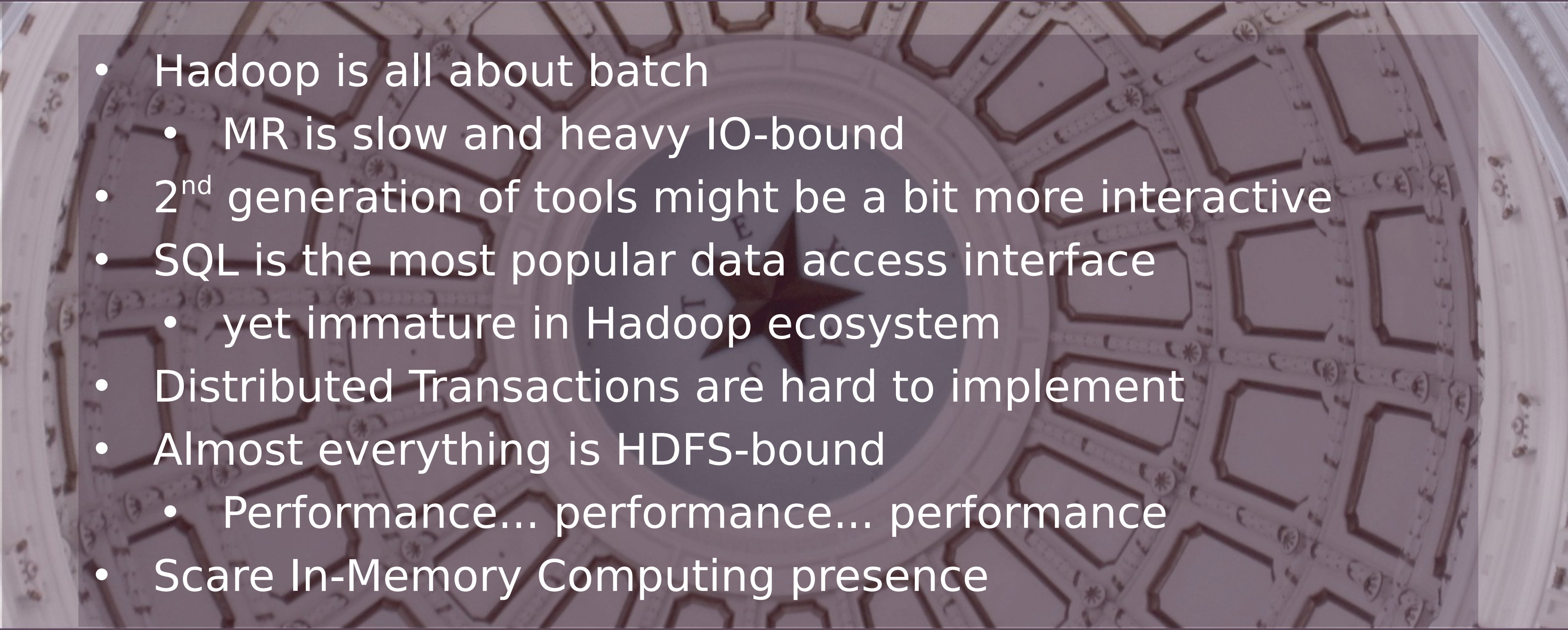
Apache Bigdata stack

- Bigtop is the cutting edge of Apache Bigdata stack
- Delivers:
  - A pre-cut data processing stack
  - Dev. Env. For anyone to create their own
  - Framework for easy integration/deployment/validation
    - “It works on my laptop” isn't cool anymore
- 0.x release series was focused on Hadoop ecosystem
  - Sorry, that's what we had...

10K view of Bigdata

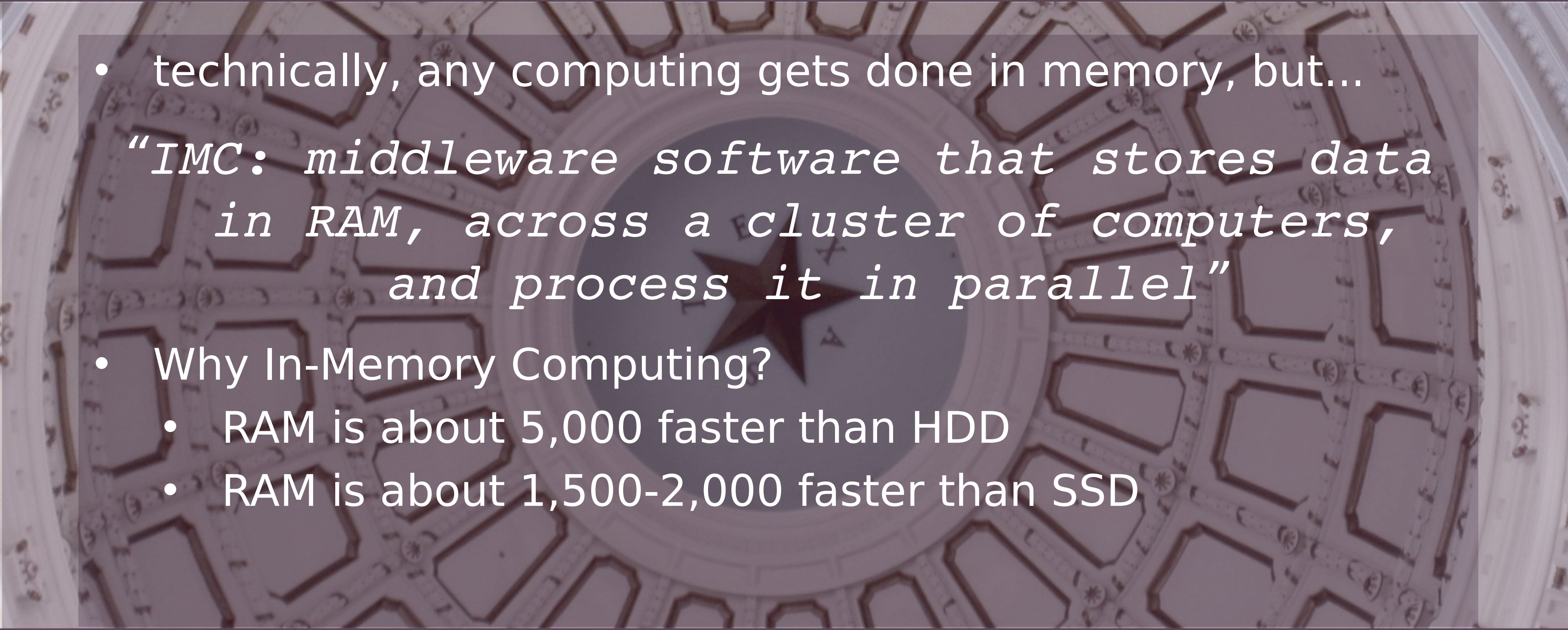
- There's more than just Hadoop (I'm shocked!)
- Hadoop is mere 5-10% of all Bigdata usecases
  - Processing documents in parallel
  - Long running processes
  - Suboptimal resource scheduling
  - Analytics and ML
- But it is NOT ideal...

What's missing



- Hadoop is all about batch
  - MR is slow and heavy IO-bound
- 2nd generation of tools might be a bit more interactive
- SQL is the most popular data access interface
  - yet immature in Hadoop ecosystem
- Distributed Transactions are hard to implement
- Almost everything is HDFS-bound
  - Performance... performance... performance
- Scare In-Memory Computing presence

IMC: what is that?

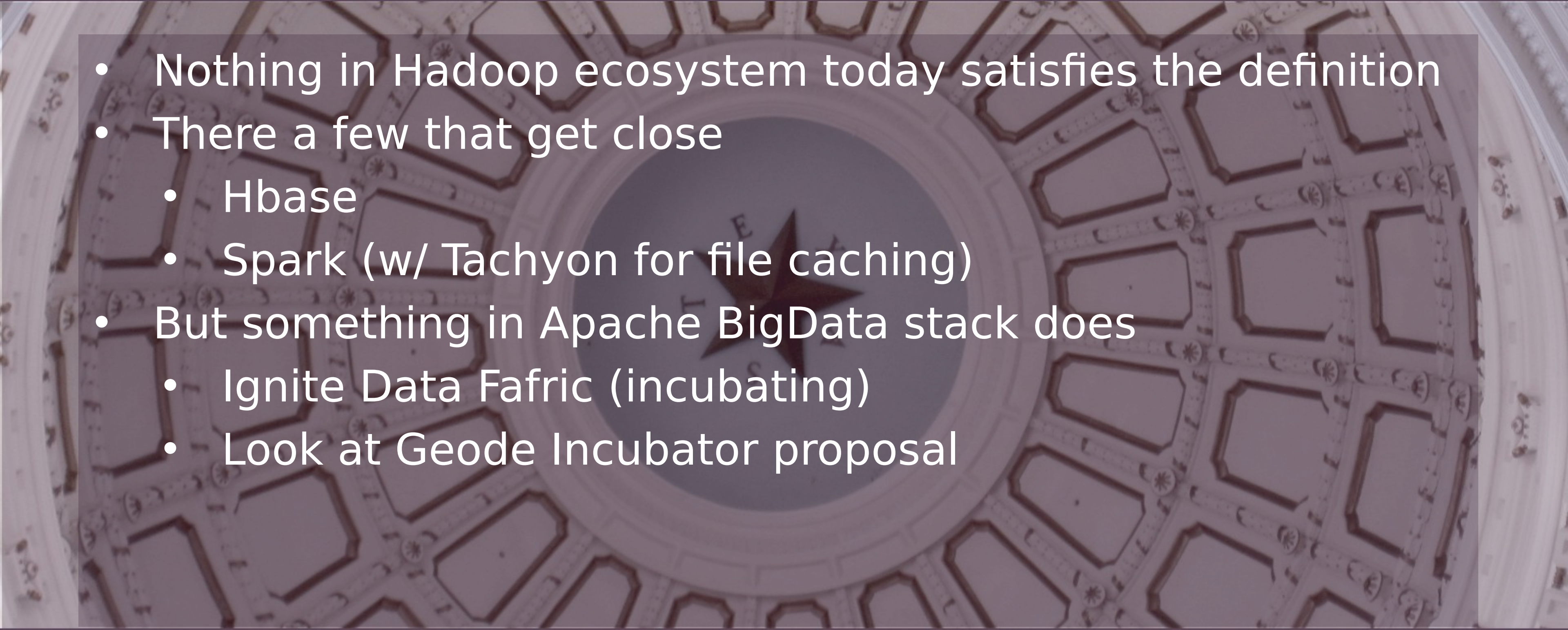


- technically, any computing gets done in memory, but…

“IMC: middleware software that stores data in RAM, across a cluster of computers, and process it in parallel”

- Why In-Memory Computing?
  - RAM is about 5,000 faster than HDD
  - RAM is about 1,500-2,000 faster than SSD

Except...

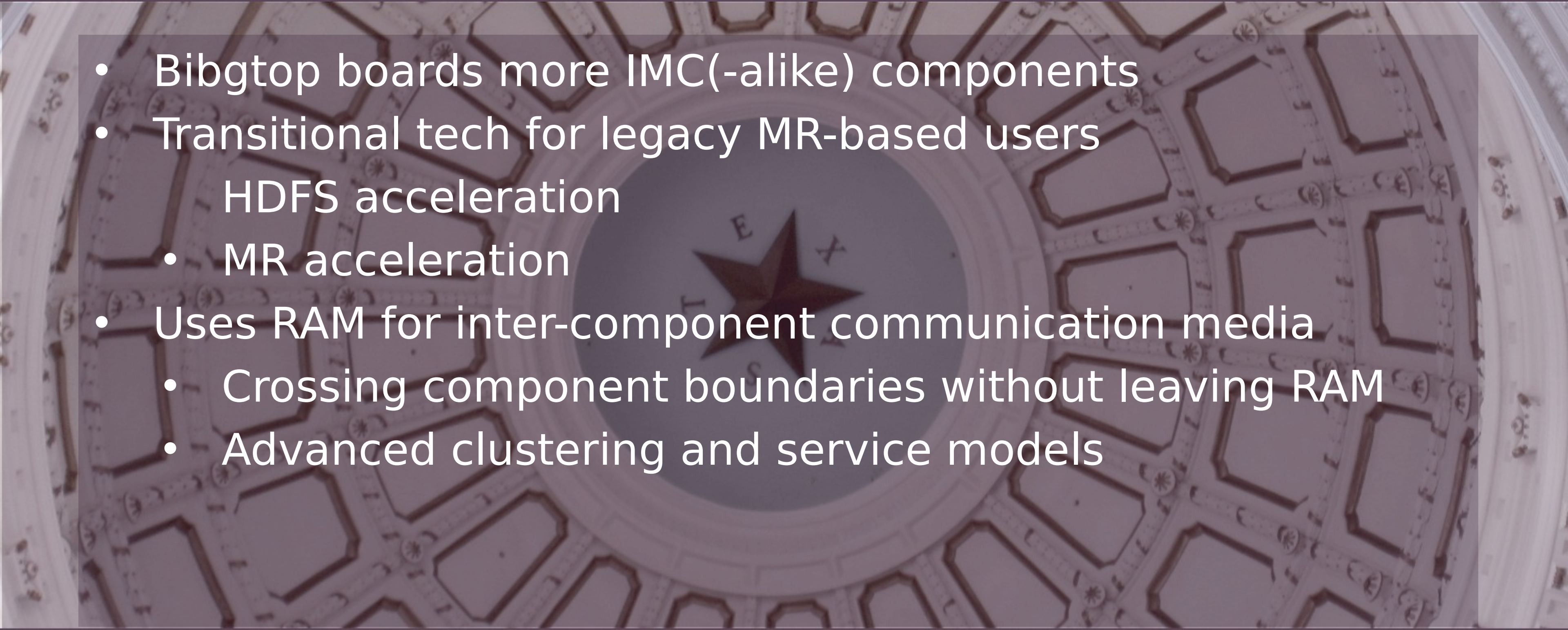


- Nothing in Hadoop ecosystem today satisfies the definition
- There a few that get close
  - Hbase
  - Spark (w/ Tachyon for file caching)
- But something in Apache BigData stack does
  - Ignite Data Fafric (incubating)
  - Look at Geode Incubator proposal

Apache In-Memory Computing #FastData

APACHECON
NORTH AMERICA

Let's get serious about IMC



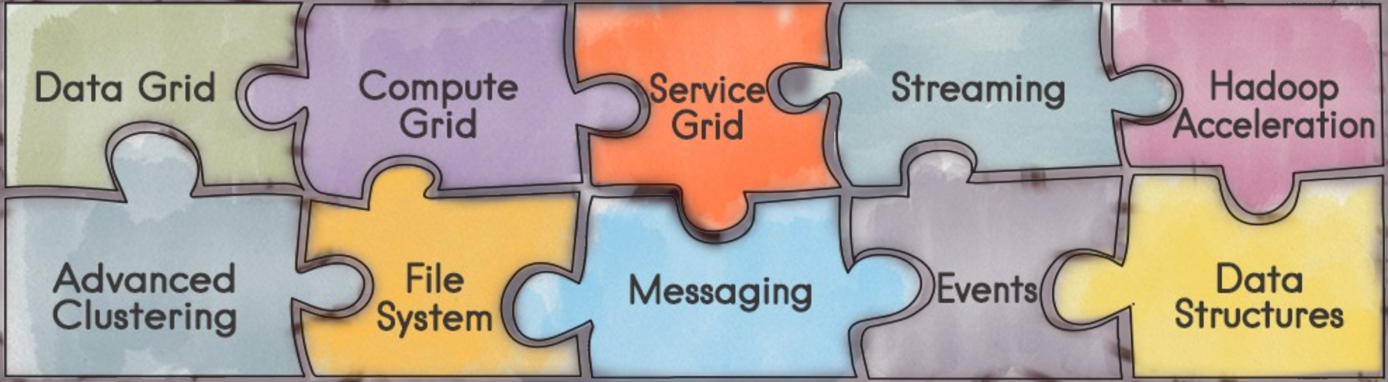
- Bibgtop boards more IMC(-alike) components
- Transitional tech for legacy MR-based users
  - HDFS acceleration
  - MR acceleration
- Uses RAM for inter-component communication media
  - Crossing component boundaries without leaving RAM
  - Advanced clustering and service models

Connecting the stack

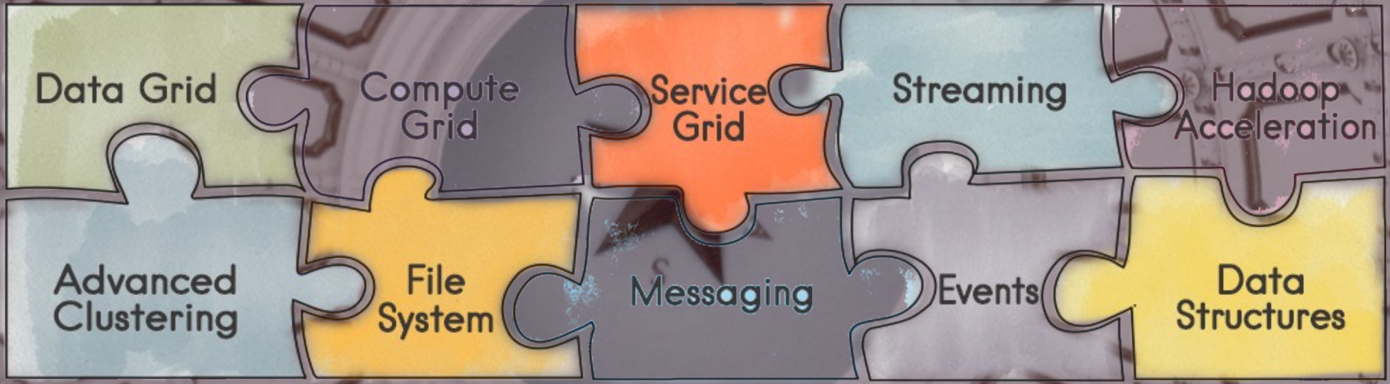
- Bigtop Data Fabric Core (Apache Ignite):
  - Works with HDFS/RDBMS/MR/Hive/Hbase/Spark/Storm/SQL
- Cluster memory is a natural media to exchange data
- Kafka --> Data Fabric RAM --> HBase --> Data Fabric RAM --> SQL querying --> Spark --> Service Singlethon --> Data Fabric RAM --> RDBMS or FS

Data Fabric: what is that?

APACHECON NORTH AMERICA

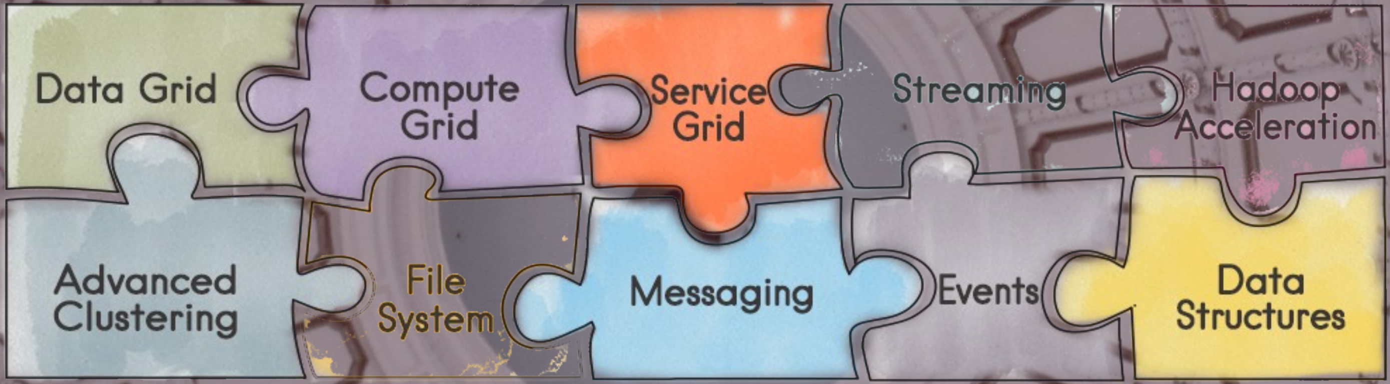


Data Fabric: customize

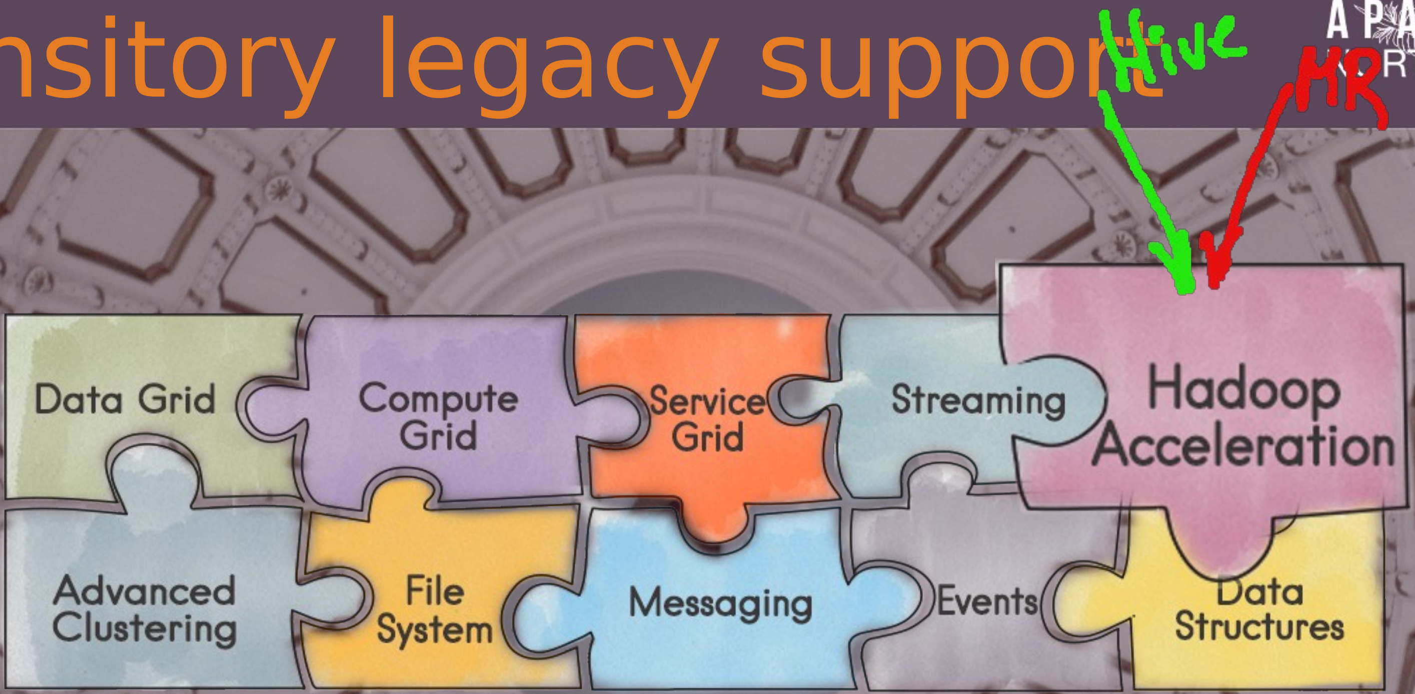


Data Fabric: … some more

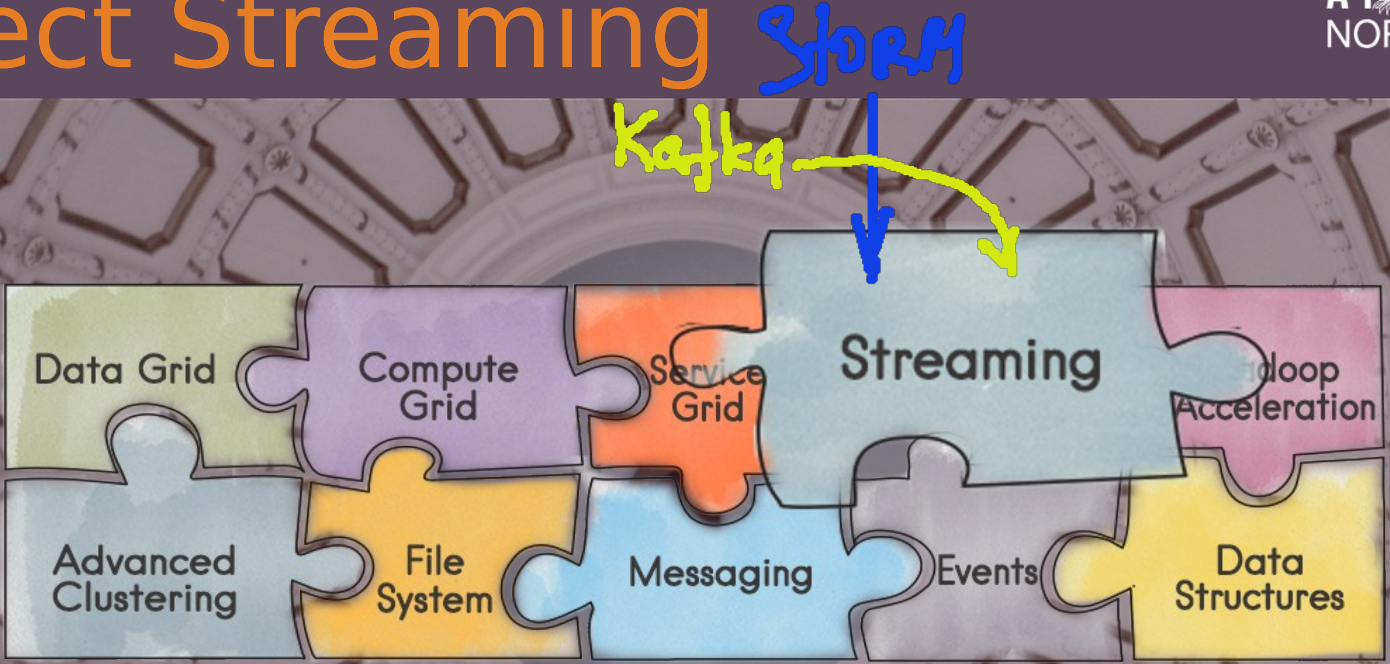
APACHECON NORTH AMERICA



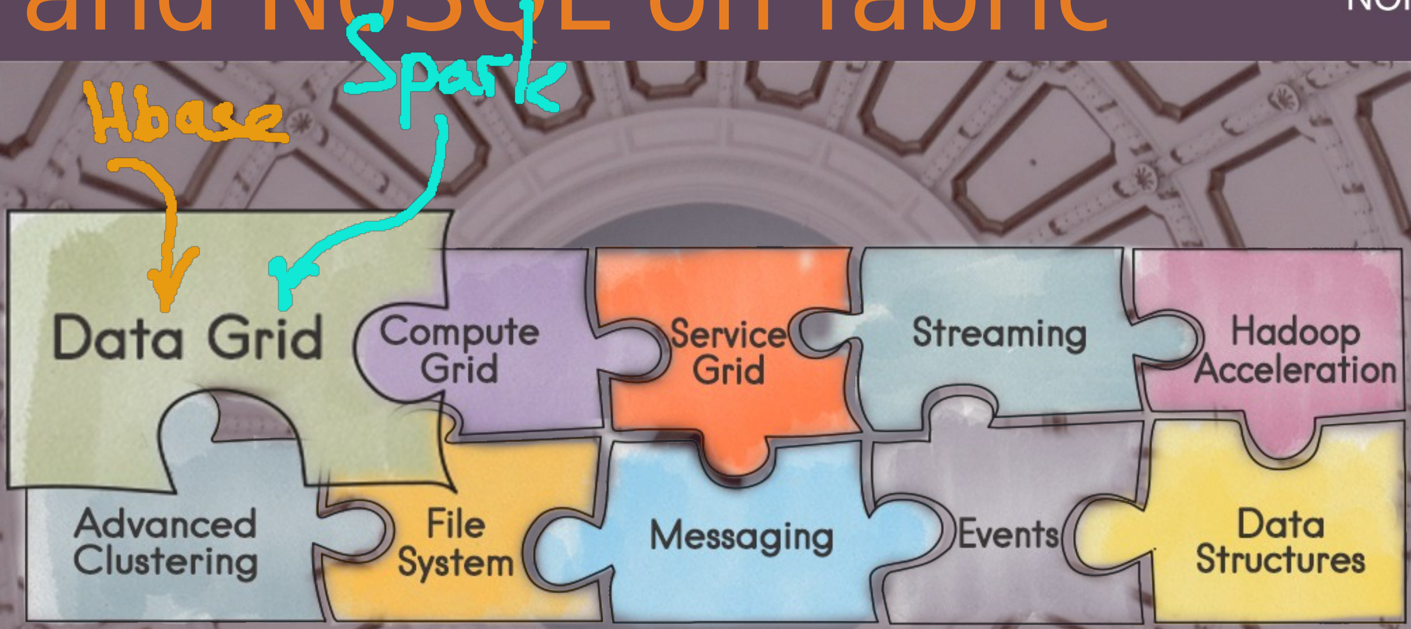
Transitory legacy support



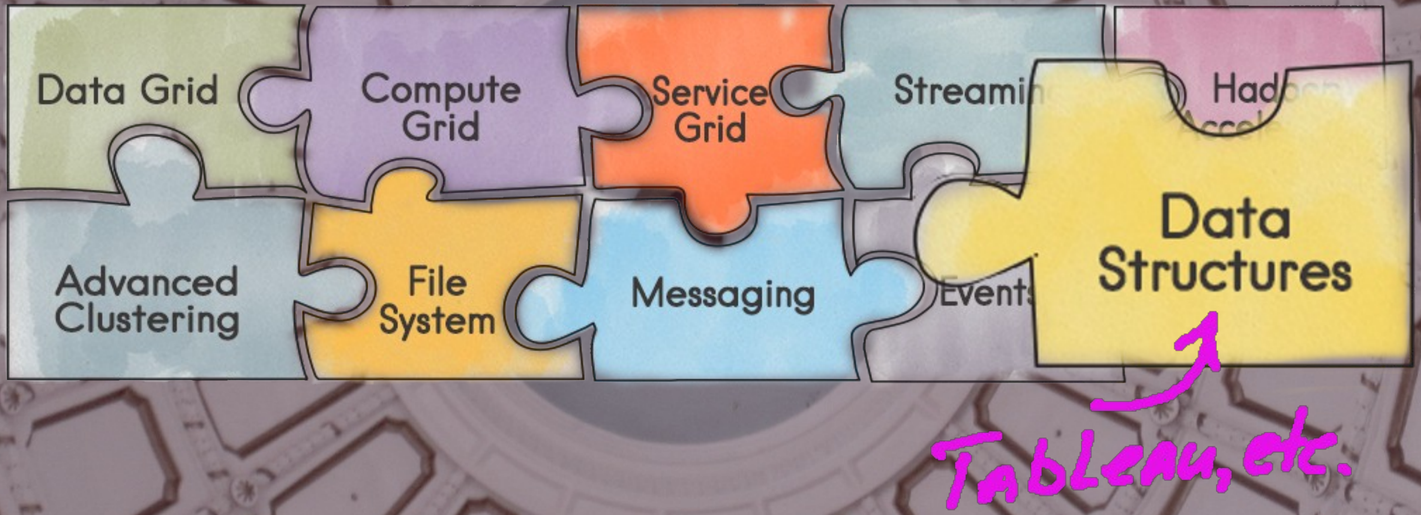
Direct Streaming



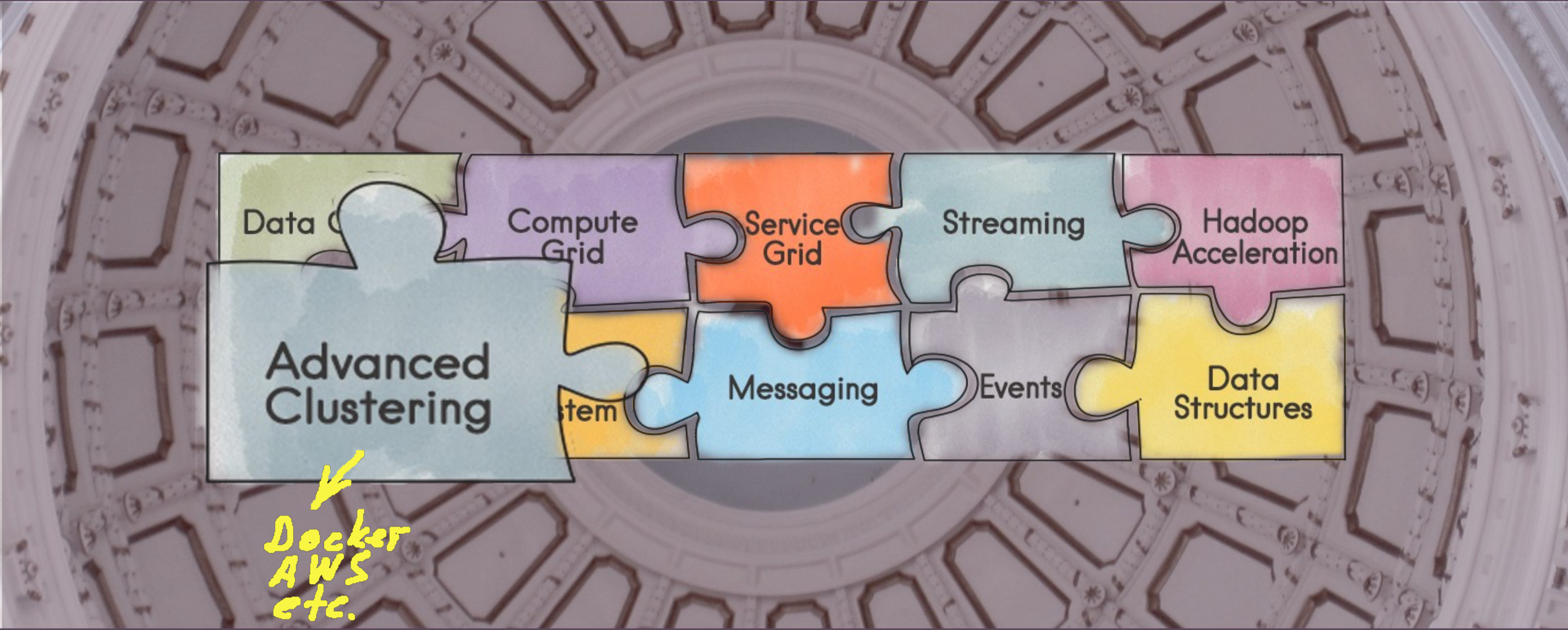
ML and NoSQL on fabric



Analysing w/ 3rd party tools

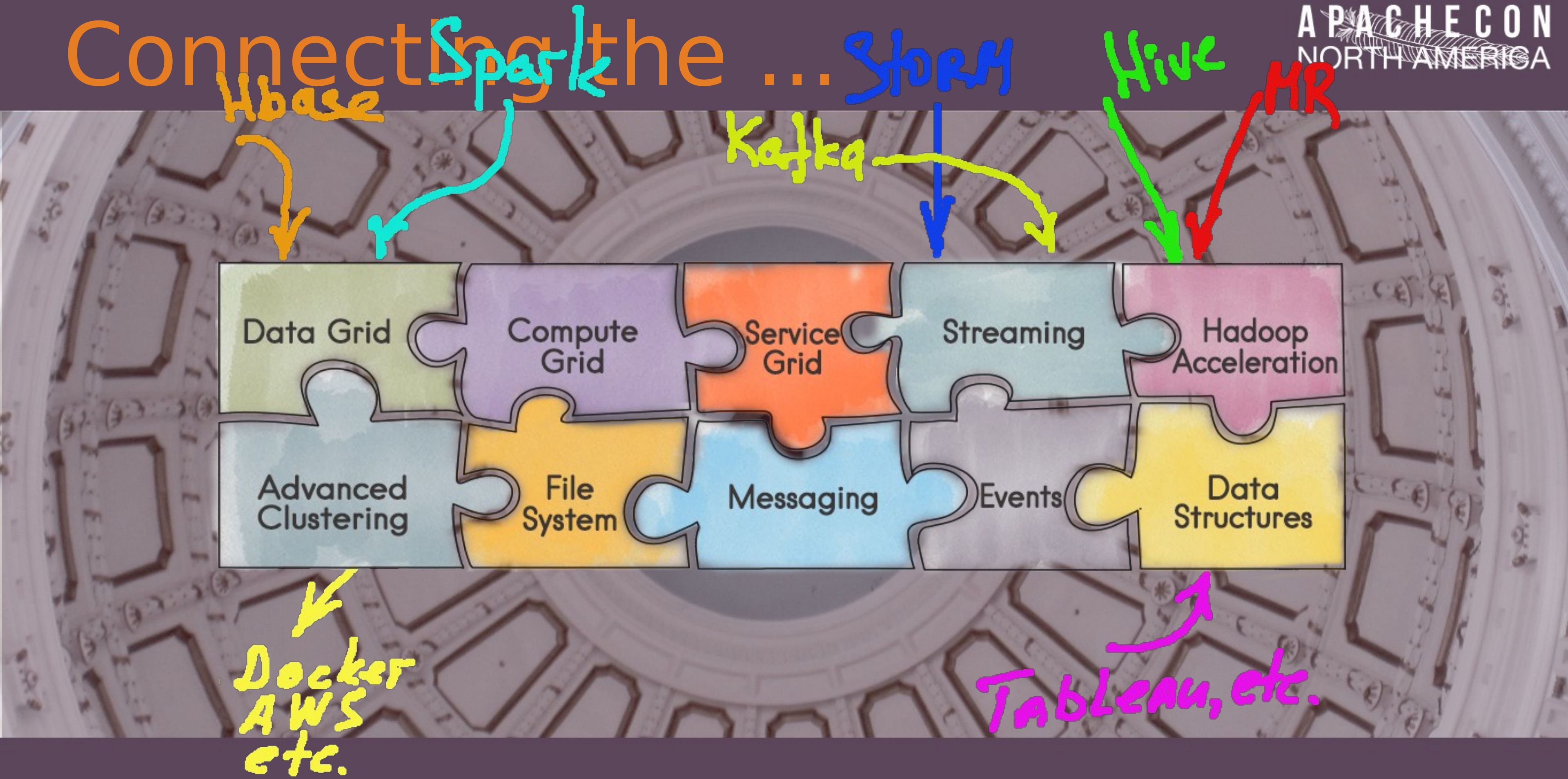


Deploy nodes everywhere

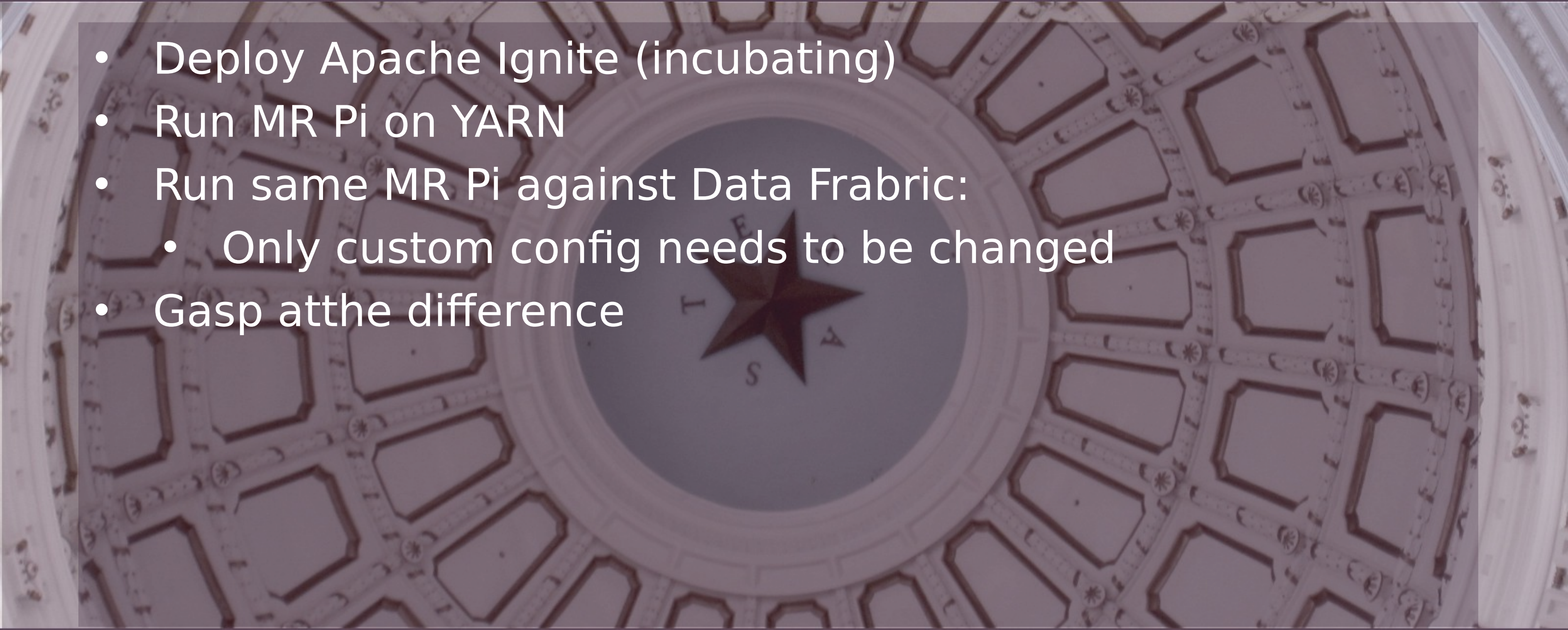


Docker
AWS
etc.

Connecting the ...



Live Demo

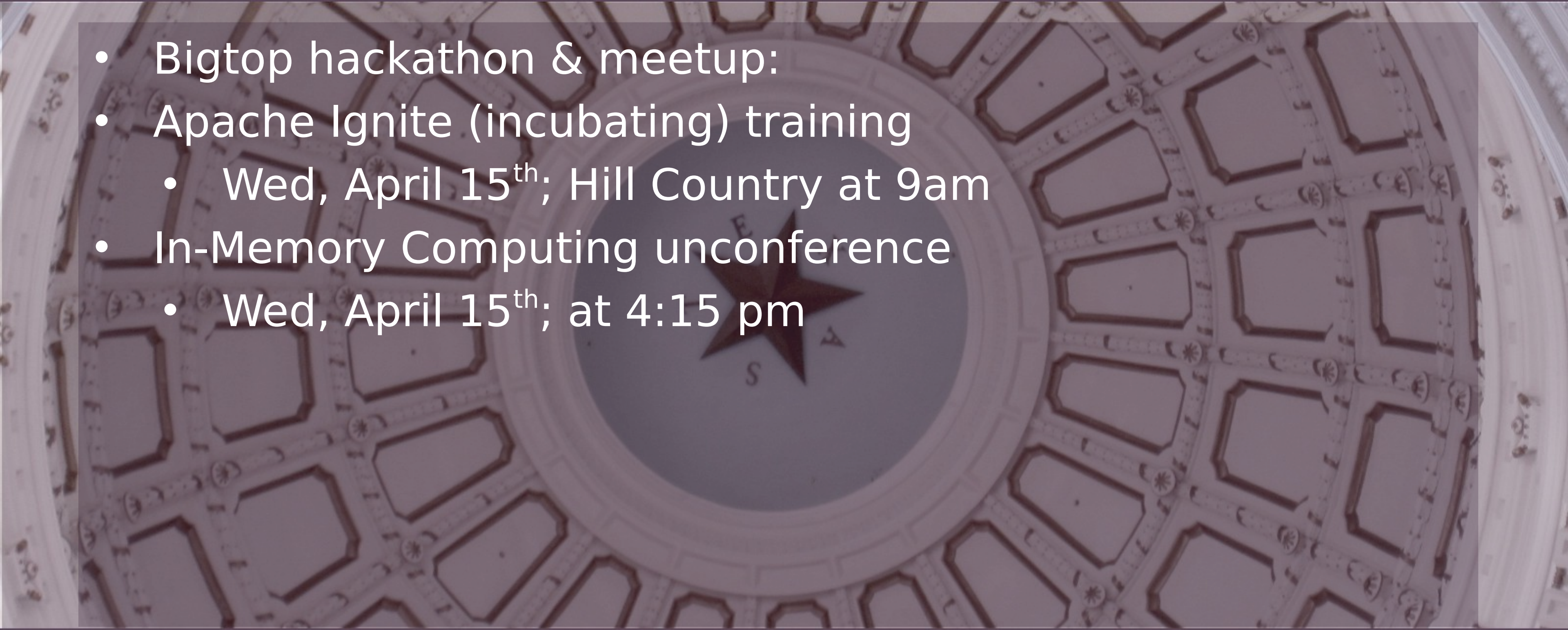


- Deploy Apache Ignite (incubating)
- Run MR Pi on YARN
- Run same MR Pi against Data Frabric:
  - Only custom config needs to be changed
- Gasp atthe difference

mapred-site for IMC MR

```
<configuration>
    <property>
        <name>mapreduce.framework.name</name>
        <value>ignite</value>
    </property>
    <property>
        <name>mapreduce.jobtracker.address</name>
        <value>localhost:11211</value>
    </property>
    <!-- Parameters for job tuning. -->
    <!--
              mapreduce.job.reduces
              mapreduce.job.maps
    -->
</configuration>
```

Q & A



- Bigtop hackathon & meetup:
- Apache Ignite (incubating) training
  - Wed, April 15th; Hill Country at 9am
- In-Memory Computing unconference
  - Wed, April 15th; at 4:15 pm

Dr. Konstantin Boudnik

@c0sin

cos@apache.org

APACHECON NORTH AMERICA