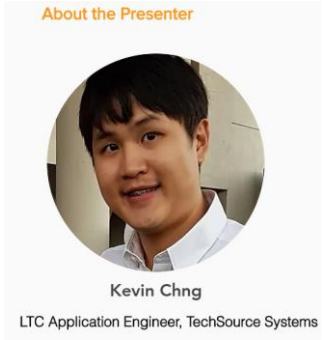
TECHSOURCE
Dynamic Solutions Precise Results

Technical Discussion: Big Data From SQL Database



Kevin Chng
LTC Application Engineer
2 Years with TechSource System
(AI, Finance, App Design, IT Instructure/Implementation)
Top 50 MATLAB File Exchange Contributor Worldwide
Email : kevin.chng@techsource-asia.com
Contact : +65 6842 4222

R2020a

R2019b

Supported by

MathWorks®

Southeast Asia's sole distributor of

MATLAB® & SIMULINK®

Overview

1 Access Data from SQL

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)

2 Process Big Data

Purpose-built capabilities for domain experts to work with big data

Tall Arrays

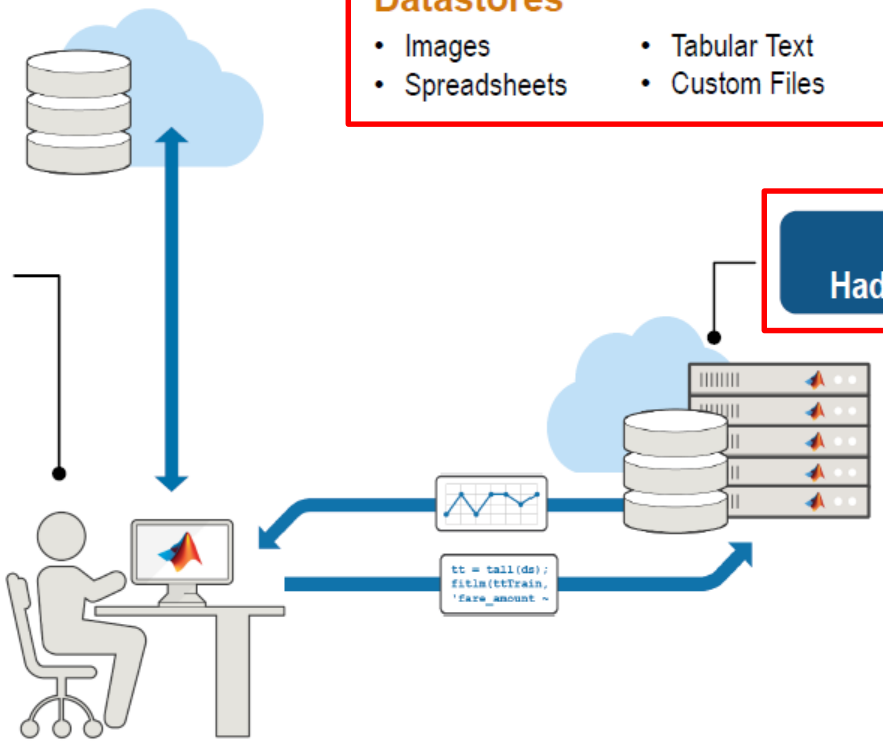
- Math
- Statistics
- Visualization
- Machine Learning

GPU Arrays

- Matrix Math
- Image Processing

Deep Learning

- Image Classification



3 Computing Resource

Scale to compute clusters and Hadoop/Spark for data stored in HDFS

Tall Arrays

- Math, Stats, Machine Learning on Spark

Distributed Arrays

- Matrix Math on Compute Clusters

MDCS for EC2

- Cloud-based Compute Cluster

MapReduce

MATLAB API for Spark

Agenda

- Challenges in handling Big Data
- Introduction to Database Toolbox, Cloud Database
- Big Data Solution in MATLAB (Datastore + Tall Array)
- New Release R2020a
- Resources & Follow-up Supports

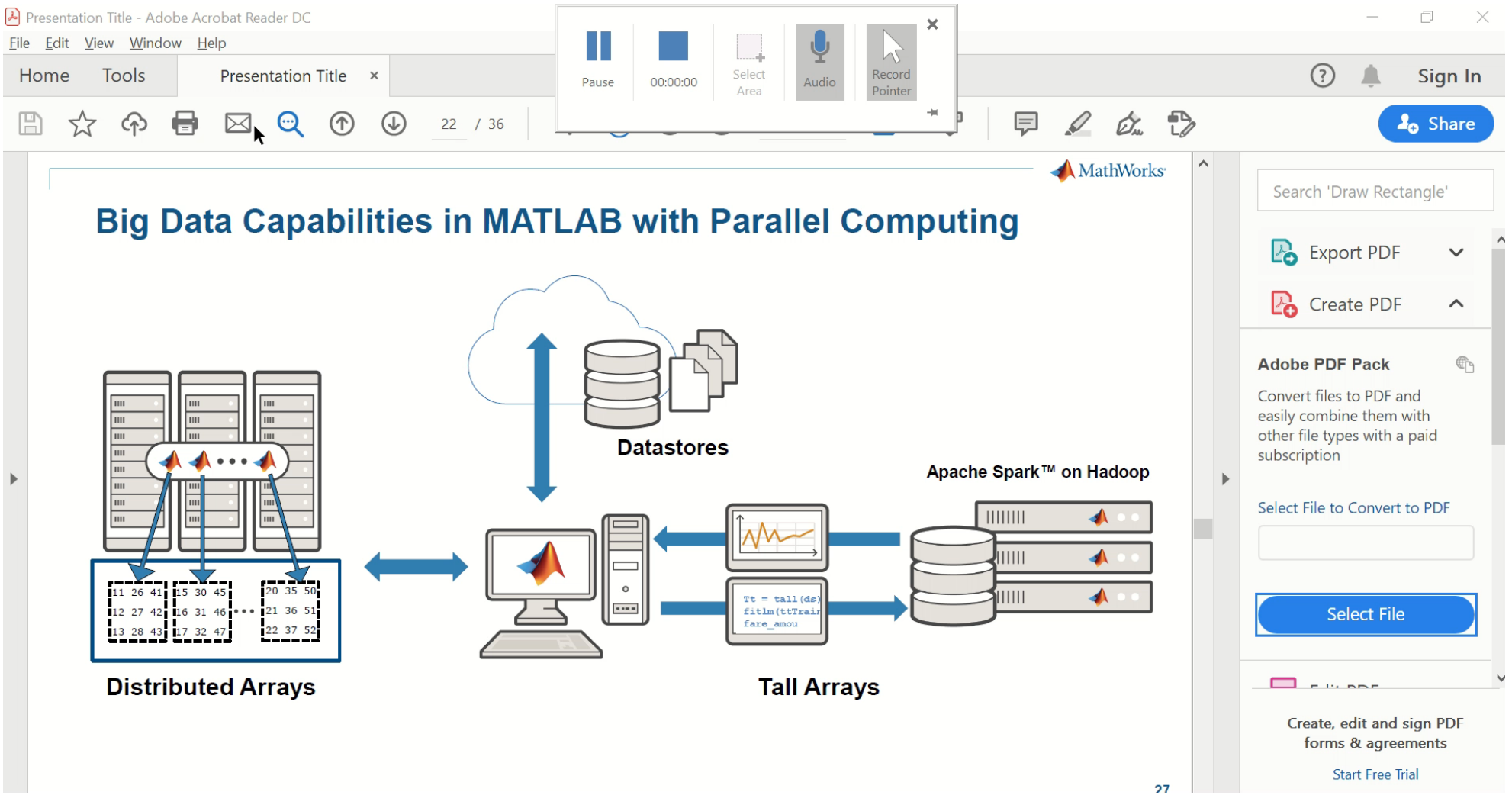
Challenges

- The Uncertainty of Data Management
- Talent Gap in Big Data
- Getting Data into Big Data Structure
- Syncing Across Data Sources
- Extracting Information from the Big Data Integration
- Miscellaneous Challenge

Introduction to Database Toolbox

- Introduction : https://www.mathworks.com/products/database.html
- Choosing Between ODBC and JDBC Drivers : https://www.mathworks.com/help/database/ug/choosing-between-odbc-and-jdbc-drivers.html
- Configuring Driver and Data Source : https://www.mathworks.com/help/database/ug/configuring-driver-and-data-source.html
- Database Explorer App : https://www.mathworks.com/videos/using-the-database-explorer-app-71881.html

Example : Connect MATLAB to Microsoft Access SQL



Cloud Database

Connect to Cloud Services

Access storage, databases, and other cloud services on AWS® and Azure® from your MATLAB code.

AWS Services

- Amazon S3™
- Amazon Aurora®
- Amazon RDS (PostreSQL, MySQL®, MariaDB®, Oracle®, and SQL Server® engines)

Azure Services

- Azure Blob Storage
- Azure SQL Database
- Azure Database (MySQL and PostgreSQL engines)
- Azure SQL Data Warehouse
- Azure SQL Server Stretch Database

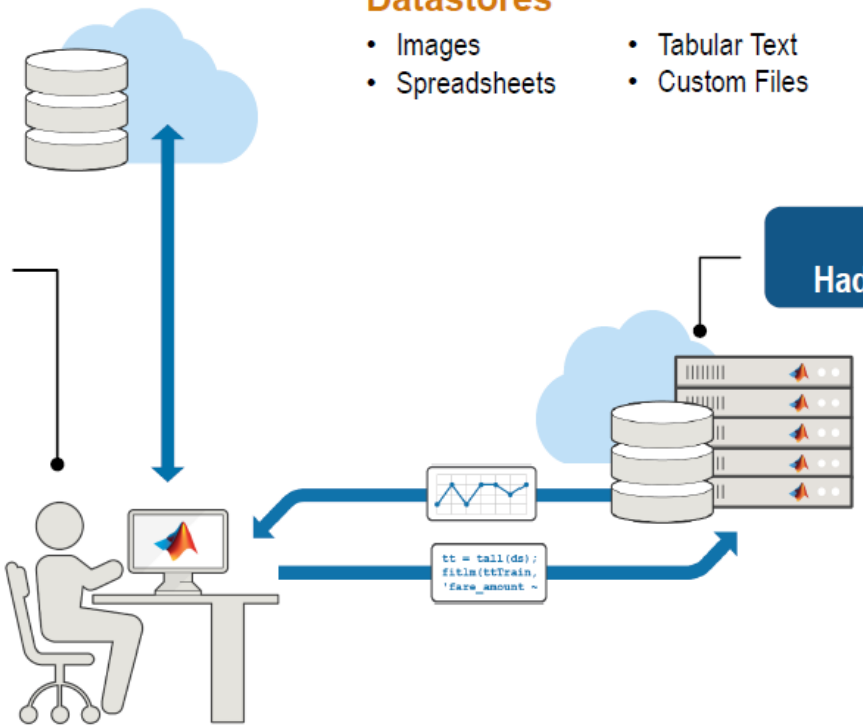
Detail : https://se.mathworks.com/solutions/cloud.html

Big Data Solution in MATLAB

Access data and collections of files that do not fit in memory

Datastores

- Images
- Spreadsheets
- Tabular Text
- Custom Files
- SQL
- Hadoop (HDFS)



Purpose-built capabilities for domain experts to work with big data

Tall Arrays

- Math
- Statistics
- Visualization
- Machine Learning

GPU Arrays

- Matrix Math
- Image Processing

Deep Learning

- Image Classification

Scale to compute clusters and Hadoop/Spark for data stored in HDFS

Tall Arrays

- Math, Stats, Machine Learning on Spark

Distributed Arrays

- Matrix Math on Compute Clusters

MDCS for EC2

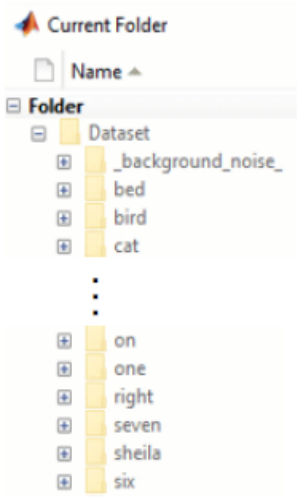
- Cloud-based Compute Cluster

MapReduce
MATLAB API for Spark

Big Data Solution in MATLAB

Datastores

- You don't need to load all your image/signal data into memory.
- Loads in data only when needed

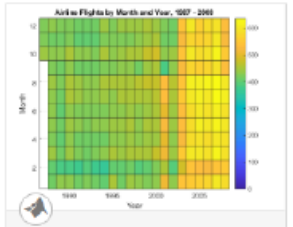


`datastore`
`tabularTextDatastore`
`spreadsheetDatastore`
`imageDatastore`
`parquetDatastore`
`fileDatastore`

Custom Datastores also available

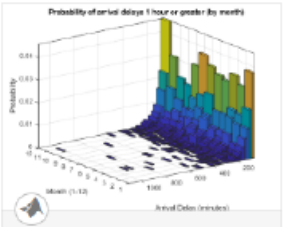
Tall Arrays

- Work with out-of-memory numeric data.
  - Train deep neural networks for numeric arrays



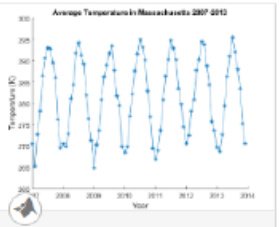
Analyze Big Data in MATLAB Using Tall Arrays

Use tall arrays to work with big data in MATLAB®. You can use tall arrays to perform a variety of calculations on different types of



Histograms of Tall Arrays

Use histogram and histogram2 to analyze and visualize data contained in a tall array.



Process Big Data in the Cloud

Access a large data set in the cloud and process it in a cloud cluster using MATLAB capabilities for big data.

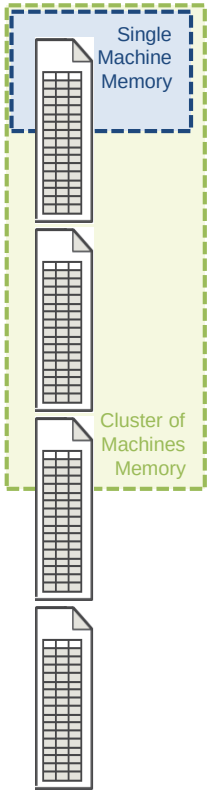
Pre-processing Big Data

tall arrays in R2016b

- New data type designed for data that doesn't fit into memory
- Lots of observations (hence "tall")
- Looks like a normal MATLAB array
  - Supports numeric types, tables, datetimes, strings, etc…
  - Supports several hundred functions for basic math, states, indexing,etc.
  - Statistics and Machine Learning Toolbox support (clustering, classification,etc)

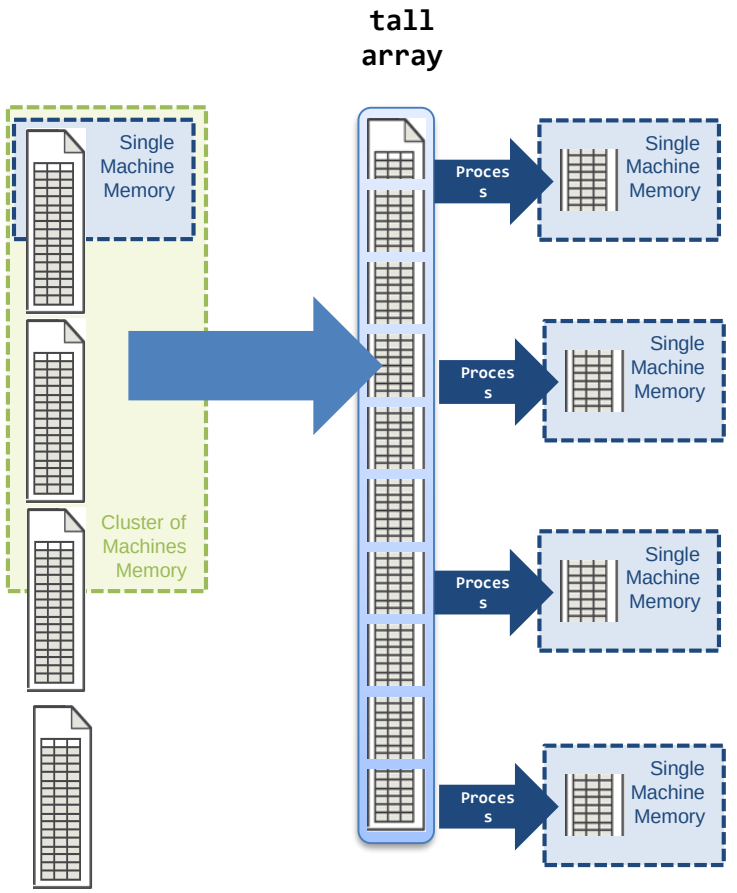
tall arrays in R2016b

- Data is in one or more files
- Typically tabular data
- Files stacked vertically



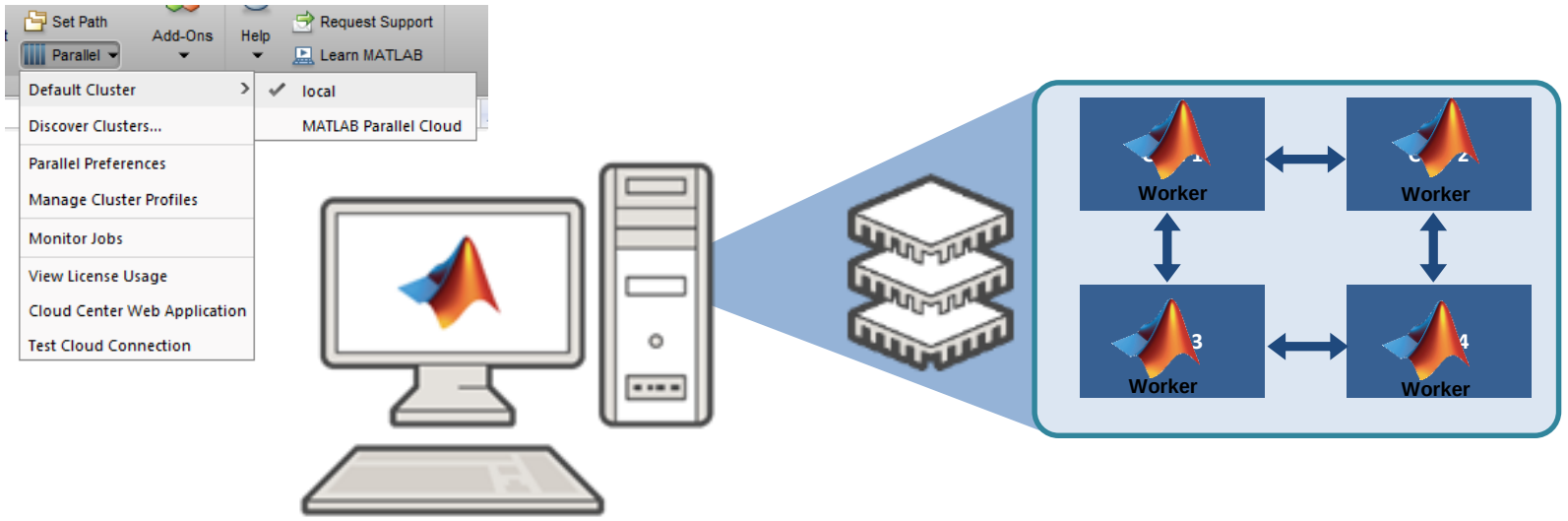
tall arrays in R2016b

- Automatically breaks data up into small “chunks” that fit in memory
- “Chunk” processing is handled automatically
- Processing code for tall arrays is the same as ordinary arrays



Computing Resource in your PC

Multicore Desktop

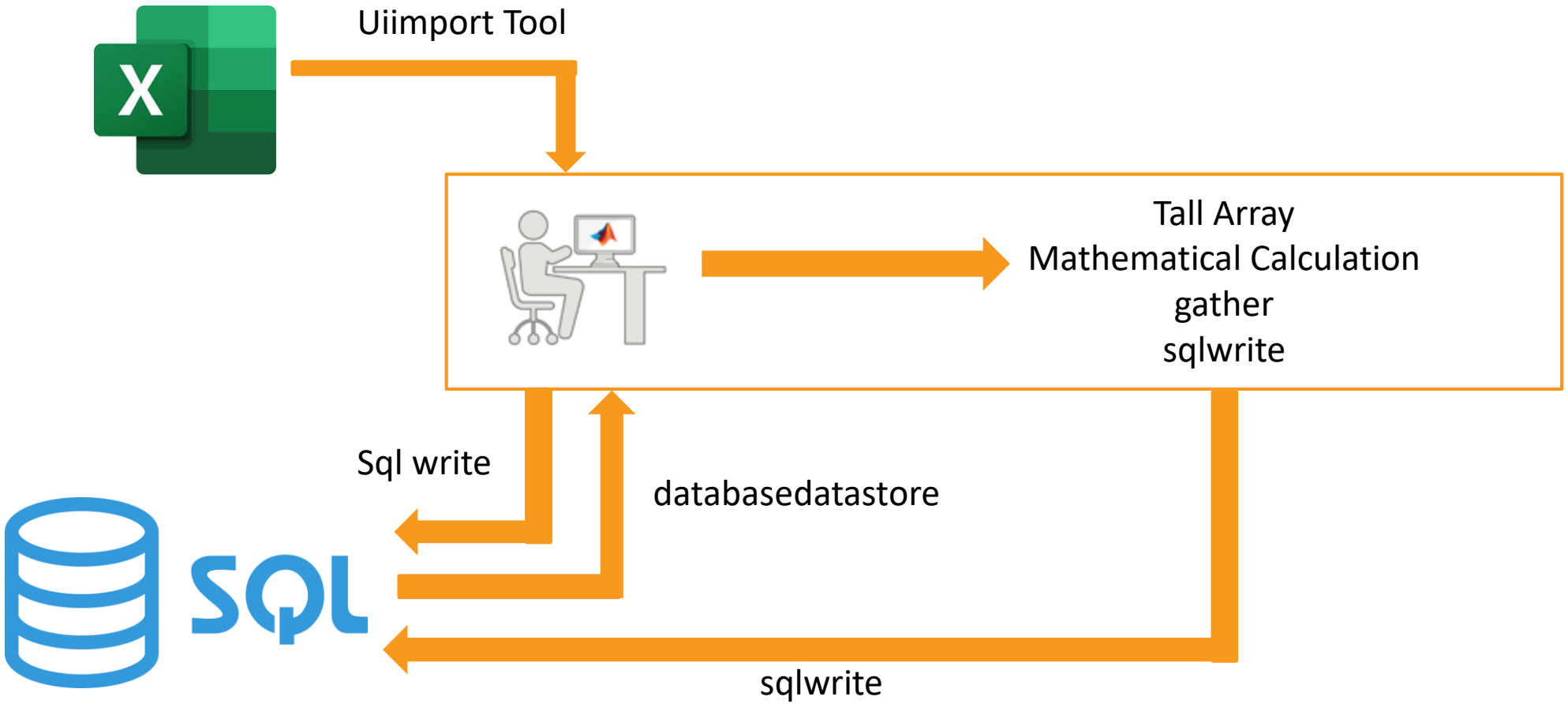


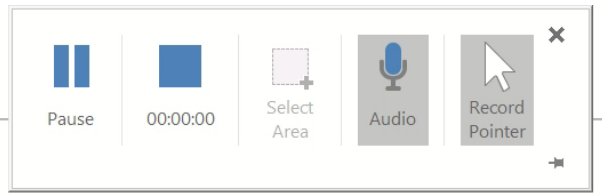
Computing Resource in Cluster : MATLAB Parallel Server

Cluster Hardware/Server

Example Big Data processing for SQL database

R2019b





Example Big Data processing for SQL database

R2019b

New release 2020a

- R2020a at Glance : https://www.mathworks.com/products/new_products/latest_features.html

- MATLAB Update : https://se.mathworks.com/help/matlab/release-notes.html
  - Live Editor (execution speed, more GUI)
  - Tall Array
  - Datastore (writeall function)

Resources & Follow Up

- Documentation : https://www.mathworks.com/help/database/import-large-data-programmatically.html
- Local AE Supports (TechSource System)
- Specialized Applications : Global AE Supports (MathWorks) : Simon Ng (Contact Point)
- R2020a at Glance : https://www.mathworks.com/products/new_products/latest_features.html

TECHSOURCE
Dynamic Solutions Precise Results

THANK YOU

Kevin.chng@techsource-asia.com

Southeast Asia's sole distributor of
MATLAB® & SIMULINK®

Dynamic Solutions. Precise Results.