

Proof of Technology



Everything you need to know about hosting a

a local Spark Event

Table of Contents

[Foreword 2](#_Toc452037158)

[Overview 3](#_Toc452037159)

[Step 1 Requesting 3](#_Toc452037160)

[Step 2 Scheduling / Resource Allocation 4](#_Toc452037161)

[Step 3 Preparation 4](#_Toc452037162)

[Hosting Options 4](#_Toc452037163)

[BlueMix (SaaS) 5](#_Toc452037164)

[Data Scientist Workbench (DSWB) 5](#_Toc452037165)

[IBM Supplied Equipment 5](#_Toc452037166)

[Registration 6](#_Toc452037167)

[Content 7](#_Toc452037168)

[Venue 7](#_Toc452037169)

[Training Rooms and Furnishings 7](#_Toc452037170)

[Materials 8](#_Toc452037171)

[Survey (need input from Braden) 9](#_Toc452037172)

[Roles and Responsibilities 9](#_Toc452037173)

[Local ACA 9](#_Toc452037174)

[North American Spark Team 10](#_Toc452037175)

[Step 4 Event Day 10](#_Toc452037176)

[Room Prep 11](#_Toc452037177)

[Delivery 11](#_Toc452037178)

[Survey 11](#_Toc452037179)

[Assistance 11](#_Toc452037180)

[Clean Up 12](#_Toc452037181)

[Step 5 De-Briefing 12](#_Toc452037182)

[Determine call to action?? 12](#_Toc452037183)

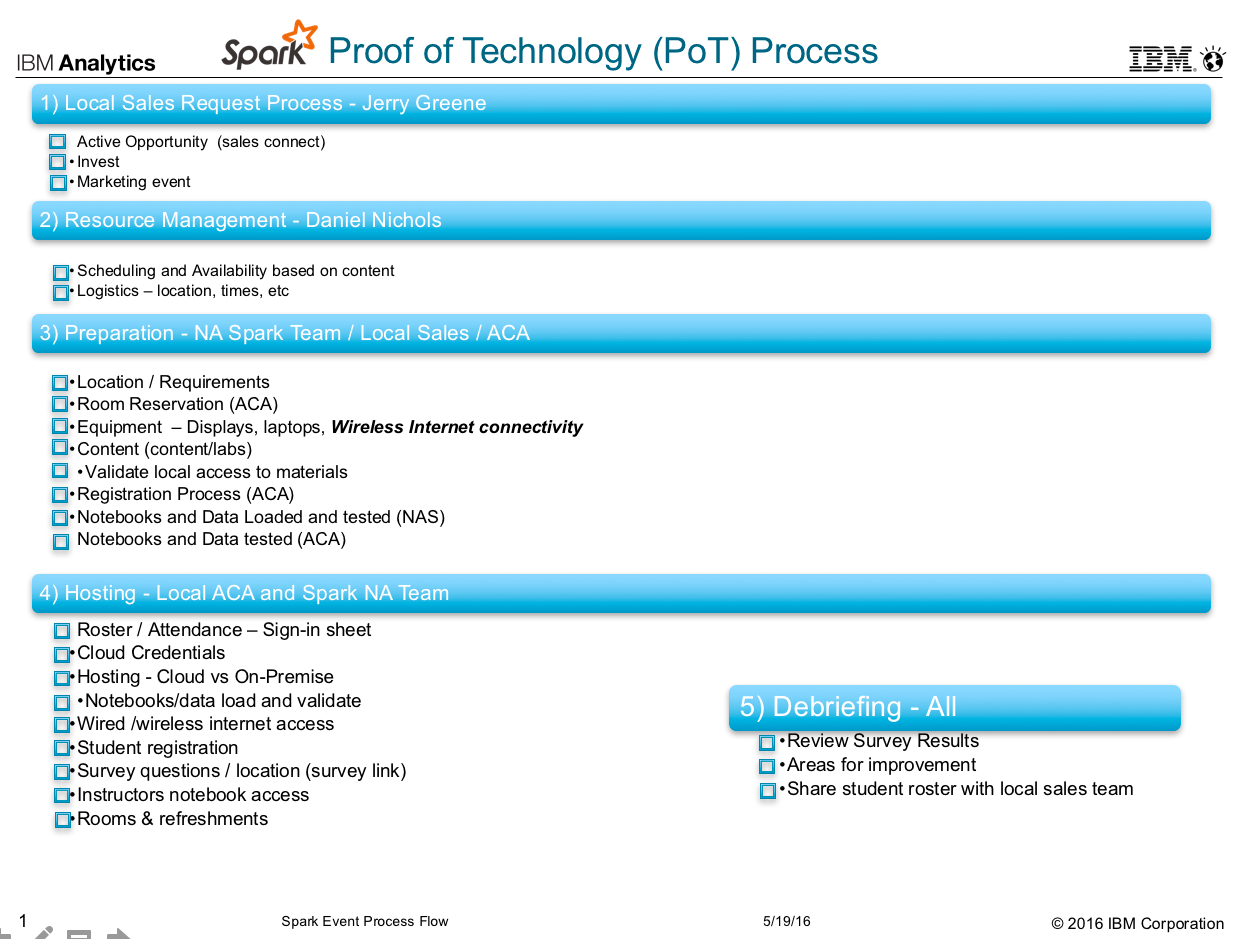
[Goodies 13](#_Toc452037184)

# Foreword

The target audience of this document is any in region or above region resources interested in leveraging the North American Spark (NAS) team lead by Dan Nichols in pursuit of contributing to IBM revenue, while increasing IBM’s presence and contributions to the Open Source Community. Based on interest and current scheduling the Open Source Pursuit team might also be interested in this document which could also be used as a check list for each stage.

If there is anything you think might be helpful to include please let us know.

# Overview



# Step 1 Requesting

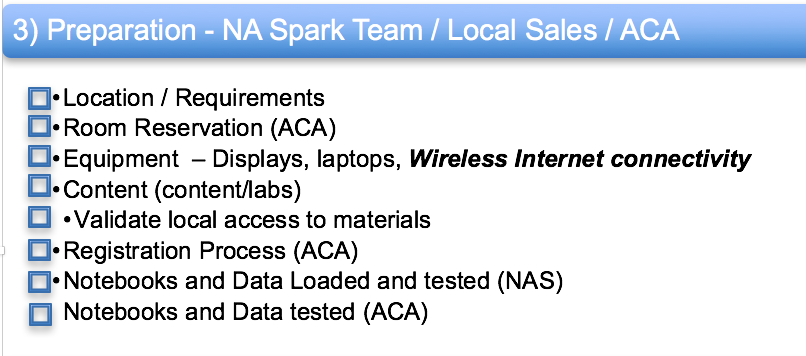
Spark PoT or Spark Days requests are being routed to Jerry Greene ( [jerry.green@us.ibm.com)](mailto:jerry.green@us.ibm.com)) to determine the type, and priority of activity ranging from immediate active opportunities, invest, and or direct (IBM sponsored) or indirect marketing events (MeetUps, DataPalooza, Strata, Summit). Please keep in mind and respect that a Sales Connect number is necessary for all active opportunities.

# Step 2 Scheduling / Resource Allocation

Daniel Nichols is running point on resource availability, allocation, and taking a stab at duration, content and time frame.

Upon scheduling he will initiate an email to Spark-NA which will notify every one of the date and location. In addition, the event will be logged in our engagement/schedule spreadsheet stored on Box.

Behind the scenes Braden will update the survey to include the new logistics in the survey drop down menu.



# Step 3 Preparation

Preparation needs to begin once a time frame and location have been established. Key contacts will be identified from both the NA Spark Team as well as the Local teams. It is really helpful if preparation begins 3-4 weeks prior to the event date to allow for all the logistics and planning. Please take a look at the diagram of what’s involved with just preparation.

The primary contact for the NAS team is Daniel Nichols (dannicho@us.ibm.com ). The instructors will be the NAS Team joining forces with the local ACA responsible for coordinating efforts with local and remote resources. They will be instrumental and helpful in determining the best hosting option of which there are several alternatives.

## Hosting Options

1. In the cloud using BlueMix (Spark as A Service)
2. In the cloud using Data Scientist Workbench (DSWB)
3. On-Premise using:
   1. An IBM supplied laptop running Windows 7 64 bit natively or via a VM
   2. An IBM supplied laptop running BigInsights 4.2

Common denominator is ***internet access*** is required for several of the options not just for hosting but also to access registration, materials, and surveys.

Please note we are striving to get others into the cloud so please lead with one of the following two options

### BlueMix (SaaS)

By far the most desirable to use environment since it’s something we sell. Some negatives include supported tools like R Studio, and we need to mention that this is what the sellers are compensated on.

Students need to request access to BlueMix Spark-as-a-Service (SaaS). This needs to be done in advance of the class. https://console.ng.bluemix.net/

### Data Scientist Workbench (DSWB)

Labs were built and delivered in Washington and Chicago using DSWB, were we experienced "Outstanding support and scalability".

Students need to request access in advance of the class where they will receive a userid and create a password. http://www.datascientistworkbench.com

### IBM Supplied Equipment

Coordinated thru the Tech Exploration centers (<http://w3-03.ibm.com/software/spcn/content/U372058Y51208T37.html)> and requires 3-6 weeks in advance for machines to configured, loaded, and tested in advance. Closest one for me on theist coast is use San Francisco, Chicago Boston, … Best way to communicate to this team is through their community <https://w3-connections.ibm.com/communities/service/html/communityview?communityUuid=fdc19109-03a4-48e0-8509-123c1c91b8f8>

If laptops are supplied, then we will have to also notify the tech of what asset needs to be loaded or provide them with the image to load. There are a few choices and each has its own assets and liabilities

#### Alternative 1 Physical Windows 7 Machine

In the larger cities if we host the event at an IBM Technology center like DC, Chicago, Boston, … the class rooms are equipped with windows machines that already connect to the internet. If classrooms do not have machines they can also be ordered from a technology center

#### Alternative 2 Virtualized Windows 7 Machine with Internet Connectivity

#### 

Equipment and the ability to run a virtualized machine can be ordered from any IBM Technology Center. The Equipment will be shipped to specified location where it will be used and then returned to IBM. When talking to the IBM Tech please let them know you are looking for the Spark PoT Image.

#### Alternative 3 Windows 7 professional (native or VM) with a Linux VM

If VMware is used require it be running with VMWARE 11 and the BigInsights VM or cloud image is configured as a single or multi node cluster running BI Cluster 4,1

Need to choose a version – quick start. Gb/Dan K to work with the tech to know which asset to use. I really like the most recent image Dan K supplied and will need to use that. This was we can gain expertise by upgrading Spark 1.6 to 2.0 (WIP)

Will need to route the Aspera image into the techex so that it can be catalogued and provisioned once its fully loaded and tested with the PoT (WIP)

#### Alternative 4 Customer Supplied laptop

In many cases we encourage students to bring their machine provide the have access to wireless network with internet access. Customer supplied machines carry some restrictions and potential hazards depending on the amount of local disk space, ram, age and number of processors, firewall configuration, or other network security issues.

While it’s not recommended it could work provided they have VMware Player installed and running.

Don’t forget since the images are using licensed Windows and Red Hat copies they

would need to be copied, used, and deleted at the end of the day as they are both licensed copies of Windows and Red Hat to protect IBM.

## Registration

Please be aware we are offering two different events a Spark Day and a Spark Proof of Technology (PoT). Spark Days request process is managed by Jerry Green while the

Dan Nichols is managing the PoT.

Spark PoT’s leverage the Eventbrite invitation / registration process and is a closed event to only those who have registered. Once an event date is determined an invitation and link will be generated by the Open Source Analytics Pursuit Team and shared.

Registration being tracked by open source.

## Content

The PoT consists of both a lecture and lab which can be done in a variety of environments. Within those environments there are labs, and data sets used to illustrate functionality. In general, we have always done 2-3 labs/topics in an all-day event. We are starting to do half day events with is reducing the number of labs from three to two. Probably not practical to think that you can do more than 2-3 labs in an eight-hour session.

As of a few weeks ago our team is prepared to offer the following topics

Lab1 – Basic Data wrangling

Lab2 – Advanced Spark SQL

Lab3 – Machine Learning (recommendation)

Lab4 – Machine Learning (classification)

Lab5 – WIP – Kafka

Lab5 – WIP – Streaming

Lab7 – WIP - SparkR – BlueMix supporting pending

## Venue

IBM field offices have been used to date and requires interaction with local office personal to schedule room requirements, Wireless connectivity. Coordinating of local logistics for rooms, refreshments, events is the responsibility of the local Sales and ACA resources

## Training Rooms and Furnishings

Local in region resource, typically the ACA will work with local facilities personal to schedule class rooms. Room requirements include:

White boards / Eraser or Flipchart, Markers and Paper

Laptop Projection devices

Power strips and cords

Wireless Internet access

Depending on the size of the room audio projection

And presenting to a large audience, additional large screens or projectors may be beneficial. Personally I host a web session and then allow everyone to log in. Be careful as viewing notebooks from a web session can be tough due to issues related to fonts and formats.

While it is not required I strongly recommend some sort of provisions be made to offer Students and Staff

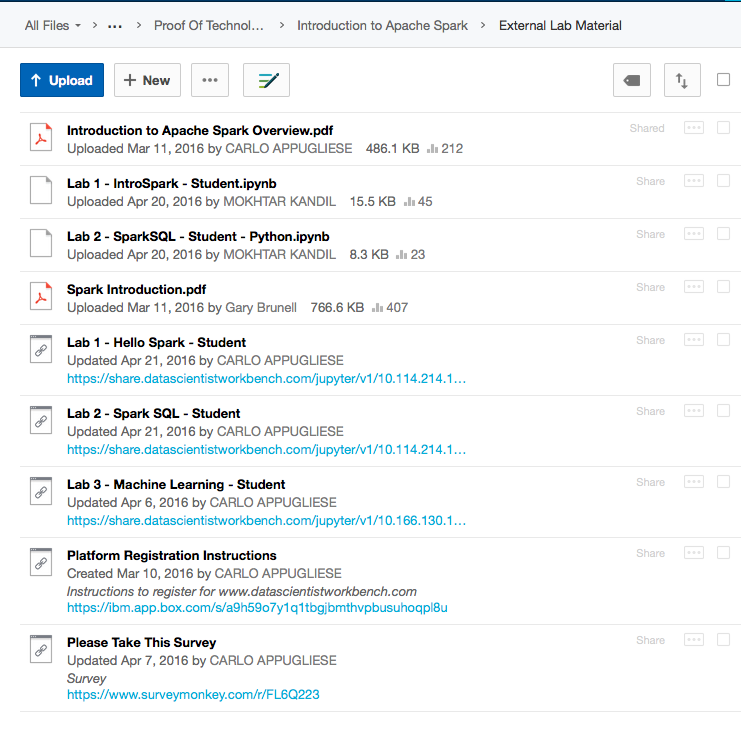
Beverages (Coffee, Soda, and water)

Snacks / Meals

## Materials

All student material is stored on box at the location https://ibm.box.com/IntroApacheSpark Password = IBMSpark please note the password is case sensitive

A Screen shot and description are listed below



**3**

**2**

**1**

1. Collateral <https://ibm.box.com/IntroApacheSpark>

Introduction to Apache Spark Overview.doc or pdf

1. Student Labs (Lab 1,2,3)
2. Completion Survey
3. Instructor Notebooks & Notebook data <https://ibm.box.com/IntroApacheSpark>-lectures

All instructor presentations and notebooks with the solution are in a different box location with a password of IBMSpark. <https://ibm.box.com/IntroApacheSpark-Lectures>

## Survey (need input from Braden)

Survey Monkey Setup

## Roles and Responsibilities

We see this process as communal one where we will divide and conquer. Representation is needed from both the NA Spark team, the in region Open Source Team, with representation from the local sales teams both technical (ACA) and non- technical resources

The Open Source Team will remain focused on proctoring and collecting questions / topics

While the Local Sales Team will remain focused on on logistics, marketing, channels like Nancy Berlin and team.

The NA Hadoop/ Spark/Open Source will remain focused on the development and delivery of the Pot presentations and labs.

For these events to be successful there must be active representation and participation from the local teams

### Local ACA

This person is the primary interface between the local teams and the North American Spark (NAS) team. Often someone with good project management skills as they need to

Determine location availaibity

Ensure room reservations have been made (currently we are averaging 3 simultaneous classes)

Ensure equipment is available (Displays, laptops, ***Wireless Internet connectivity, ...)***

Help Focus on the content/labs

Validate access to materials

Own the Registration Process (Notify Open Source Project office (Ana) to create a registration process and url. Hosts can grab flyers to circulate from the materials folder stored on box.

At this point the NAS team will begin to ensure access to notebooks, presentations, and lectures are fine. We will ask the local ACA to validate access and execution

Local ACA must drive the campaign hard locally – need to debrief with Joel and LeAnn to see what they did to coincide with a meetup????

### North American Spark Team

We are responsible for

Developing content for lectures, notebooks and their data

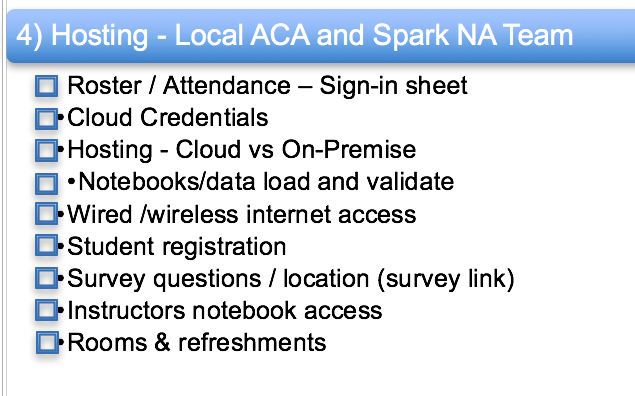
Loading and testing of notebooks and data

Initial delivery using a train the trainer approach.

All of this requires we validate wireless internet access, and notify the local ACA that the notebooks and data are ready for testing. This requires that every ACA have both a BlueMix and a Data Scientist Workbench use rid. In preparation, students need access to box <http://IBM.box.com/IntroApacheSpark> using the password of IBMSpark

We will need to decide in advance the logistics and use of tools.

* Data Scientist Workbench/Studio (DSWB) http://www.datascientistworkbench.com
* BlueMix Spark-as-a-Service (SaaS) ahead of time <https://console.ng.bluemix.net>



# Step 4 Event Day

Please print up one sign in sheet (one per classroom). A sample is included as Appendix A

## Room Prep

* Testing of laptops
* Network access (wired and or wireless)
* Environment Access (DSWB, BlueMix,)
* Notebook and data access validation
* Access to survey questions / location (survey link)
* Written on every whiteboard:

Windows userid and password

Box url and password

Instructors name and email address

Classroom number

## Delivery

Typically, we run sessions with multiple topics and multiple instructors however over time I am pretty sure this can be staffed by one or two people. When doing lab’s an ideal ratio is one proctor for every ten students.

## Survey

Can’t stress the importance of the survey as this is the only way we can tell what other topics to choose and see how we are perceived. Please be sure to let folks know that the solution link isn’t exposed till they complete the survey.

## Assistance

Some things we have seen and know how to address are

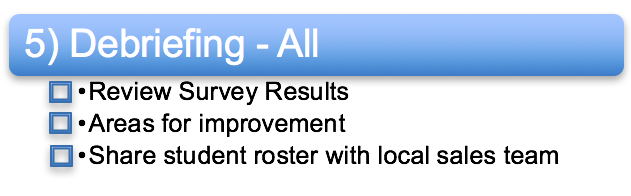
Occasionally a notebook may freeze. If so before exiting the browser, I recommend using a menu option to restart the kernel and kill outstanding jobs. A great way to copy and paste this command into your notebook

*!jps |grep SparkSubmit |awk '{print $1}'|xargs -n1 -i kill {}*

Something else to consider in advance is to notify the hosting environment support staff 2-3 weeks in advance (DSWB and or BlueMix) team. This way they are in a heightened state of alert since the issues can and should be reported directly from the notebook user interface

## Clean Up

Please be sure to help pack up the equipment and clean the white boards.



# Step 5 De-Briefing

Once survey results have been collected they will be shared and discussed on a group call. Since these sessions are gaining traction we are continually looking to see what can be done to improve the quality and increase the number of advanced topics of interest. The NAS team is always looking for addition content to develop and will continue to do this based on eh demand we see from feedback like the survey.

Often attendees develop relationships with the instructors and to ensure there are no disconnects with the local sales efforts. Student rosters and the results of the survey will be distributed to the local sales team.

## Determine call to action??

Something of interest to the team might be what is our objectives aside from education/enablement.

Some food for thought are we tracking the number or rate of BlueMix accounts as a result of preparing for or upon completion of the session.

Desire to compete in a DataPalooza like event?

Get the student to share thru social media what they have done?

Get the student to upgrade from free to pay as you go?

Get them to build something ……….

Get them to host a meetup

Get them to speak at a meetup

Get them to check out whole eco system

Thoughts Ideas, suggestion from others??????

Goodies

Sessions often stimulate folk’s appetite for information. In the spirit of 1-3-9 I like to share my journey and where I go looking for answers. Some websites I use on a regular basis is

Spark Technology Center (STC) http://www.spark.tc/

Big Data University http://bigdatauniversity.com/courses/spark-fundamentals/

Hadoop-Dev https://developer.ibm.com/hadoop

BigInsights Starter Kit https://ibm-open-platform.ibm.com/biginsights/starterkits/biginsights-starter-kit-catalog/catalog.html

Apache Spark Overview <http://spark.apache.org/docs/latest/>

DataBricks Blog https://databricks.com/blog

IBM Big Data Support <https://twitter.com/ibm_bd_support>

Spark 2.0 Update <https://databricks.com/blog/2016/05/11/apache-spark-2-0-technical-preview-easier-faster-and-smarter.html?imm_mid=0e3d2e&cmp=em-data-na-na-newsltr_20160518>

Hadoop Weekly https://hadoopweekly.com/

Appendix A - Sign in Sheet



Location City / State

Class Room#

Instructors Name

Company Student Name Email