Diplomado en "Descubrimiento del Conocimiento con Herramientas Big Data", con una duración de 184 horas presenciales.

INTRODUCCIÓN. Las tecnologías que trabajan con dispositivos que capturan datos de manera continua o discreta se incrementan día a día y por lo tanto los volúmenes que se generan sobrepasan las herramientas que realizan análisis de datos al exceder la capacidad de almacenamiento o procesamiento de los equipos de cómputo, en que ellos se podrían procesar. Además de conjuntos de datos que se publican, por las nuevas políticas de un Gobierno Abierto, en sitios conocidos como Fuentes Abiertas y que en general, son históricos de años de trabajo de organizaciones públicos o privados. Nuevas tecnologías están apareciendo, para procesar este tipo de datos y de aquí de la necesidad de los profesionales de informática de aprender a utilizarlas. Este diplomado ofrece la práctica con estos conjuntos de datos, con ejercicios y ejemplos prácticos llevados a cabos por los profesores y estudiantes del Laboratorio de Ciencia de Datos y Tecnología de Software.

OBJETIVO

Ejercitar con problemas y herramientas a profesionales que deseen trabajar con proyectos de Big Data, que incentiven el desarrollo de aplicaciones analíticas en conjuntos de datos masivos (de órdenes hasta millones de registros o cientos de variables de interés) que pueden ser públicos o privados y propensos a requerir herramientas de Big Data. Desarrollos con alta probabilidad de ser innovadoras en el dominio de la información utilizado y que dan la oportunidad de conocer y utilizar herramientas que trabajan con volúmenes de datos masivos.

PERFIL DE LOS PARTICIPANTES

- Estudiantes o profesionales de la Informática, Ciencias de la Computación o áreas afines, para un mejor aprovechamiento del Diplomado.
- Deseable estén en los últimos tres o dos semestres o cuatrimestres de su carrera.
- Deseable haber desarrollado aplicaciones y tener experiencia con sistemas administradores de bases de datos, además de haber trabajado con diferentes sistemas operativos, lenguajes de programación (como Java, PHP, Javascript, Python, entre otros) y conocimientos de almacenamiento y procesamiento distribuido.
- Deseos de enfrentar retos y aprender sobre áreas de aplicación y desarrollar innovación.
- Disponibilidad de cuatro semanas con cinco días de ocho horas diarias de asistencia, además de trabajo en casa, para la evaluación de los correspondientes módulos.
- Para la obtención del diploma se valorara la asistencia, participación, propuesta de proyecto y la presentación final del avance del proyecto.

PERFIL DE LOS INSTRUCTORES

Los instructores pertenecen al "Laboratorio de Ciencia de Datos y Tecnología de Software" del Centro de Investigación en Computación (CIC), donde se desarrolla investigación y aplicaciones informáticas para análisis, deducciones, modelado, clasificación y otras operaciones sobre grandes conjuntos de datos, para facilitar la toma de decisiones y el descubrimiento de nuevos conocimientos. Extrae información útil de bases de datos y de texto, entre otros: Análisis, Tendencias, desviaciones, situaciones relevantes y, anomalías. La actividad de investigación de este laboratorio está relacionada con: Bases de datos, Probabilidad y Estadística, Inteligencia Artificial e Ingeniería de Software. La ciencia de datos es útil e indispensable para la generación de indicadores que dan respuesta puntual y objetiva en el campo de negocios, ventas, enfermedades, sus tendencias, extrapolaciones y opiniones. ¿Qué nos quieren decir los datos? Actualmente, entre otros ejemplos, se trabaja en: ¿Cómo el contexto personal, familiar y escolar afecta el rendimiento del estudiante? Analizando para ello las bases de datos de la prueba EXCALE de la SEP, Análisis de tendencias delictivas y trayectorias seguras en la ciudad, Análisis de la variación de la demanda para la optimización oportuna de la oferta del "Sistema de bicicletas públicas compartidas de la Ciudad de México" (ECOBICI), Extracción y representación de conocimiento a partir de fuentes textuales, Visualización científica, Cubos de datos, Aceleración de consultas a BD empleando Cómputo de Alto Rendimiento (HPC), Análisis de imágenes, Seguimiento de personas, así como, Detección de acciones sospechosas en sistemas de vídeo vigilancia.

Dr. Gilberto Lorenzo Martínez Luna

Estudió su licenciatura en la Escuela Superior de Física y Matemáticas, y realizó estudios en la Sección de Computación (hoy Departamento) del CINVESTAV-IPN y el Doctorado en el Centro de Investigación en Computación del I.P.N. (CIC-IPN).

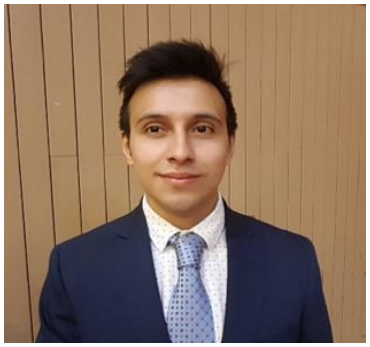
En el medio profesional ha participado en el desarrollo de sistemas de información desde 1982 en equipos de trabajo externos a compañías como Hewelett Packard en Guadalajara, Bancos como BANAMEX, el Banco Multibanco Mercantil, BANCOMER; Bolsa Mexicana de Valores S. A, Casa de Bolsa PROBURSA, Empresa Periodística Organización Editorial Mexicana S.A., Grupo consultor Merckop, empresa de Transportes Especializados en Materiales de Construcción S.A., Agencia de Publicidad Leo Burnett de México S.A., Fundación Arturo Rosenblueth A.C., además de instituciones de gobierno en que ha apoyado como asesor como en el Poder Judicial del Estado de Guerrero; o como coordinador de grupos de trabajo en SAGARPA, PEMEX – División Refinación, CFE-Gerencia de Telecomunicaciones e Informática, la Dirección General de Publicaciones de CONACULTA, la Secretaria de Salubridad y Asistencia y actualmente en investigación en el INEE aplicar Minería de Datos en Exámenes de Calidad del Logro Escolar (EXCALE, a hoy PLANEA), en problemas del Transporte de Movilidad Individual (ECOBICI), Seguridad en la CDMX (análisis de delitos y seguimiento a través de videocámaras) y proyecto de Educación para la Violencia con el CICS y SEDU de la CDMX.

En el medio docente desde 1987 ha impartido cursos a nivel Licenciatura, Maestría y Doctorado en la Universidad Autónoma Metropolitana-Azcapotzalco, en el Centro de Investigación en Computación del Instituto Politécnico Nacional, El Tecnológico de Estudios Superiores de Ecatepec, La Universidad del Valle de México-Villahermosa y la Universidad Panamericana.

El Dr. Martínez participo en la organización de eventos de Minería de Datos en el CIC-I.P.N. desde 1998 hasta 2005 y en 2015 y 2016 coordino tres Talleres cada uno de 10 días de BIG DATA con instructores del SZTAKI Institute for Computer Science and Control de Budapest, Hungría.

Desde 1996 trabaja en el CIC-IPN, y actualmente es el responsable del Laboratorio de Ciencia de Datos y Tecnología de Software donde imparte cursos; participa en la revisión de planes de Maestría; ha dirigido tesis de Doctorado(2 Graduados), Maestría(19 Graduados, 6 Actualmente) y Especialidad(5); además de proyectos de investigación (21) en sus áreas de interés como Sistemas para la Toma de Decisiones, Administración y Afinación de Sistemas Manejadores de Bases de Datos, Procesamiento Analítico en Línea, Minería de Datos, Bodegas de Datos, Visualización de la Recuperación de Información, Descubrimiento de Conocimiento en Datos Abiertos y la Estructura de la Información en Redes Complejas, y tiene cerca de 30 artículos en el área de análisis de información.

M. en C. Manuel Gutiérrez Ceballos

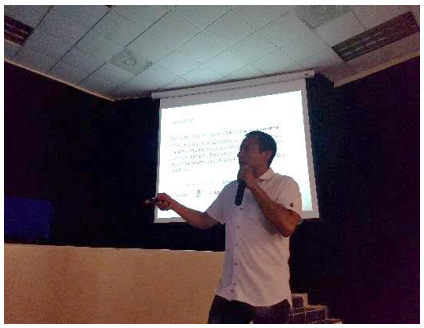


Egresado de la carrera de Ingeniería en Informática en la Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas (UPIICSA) del Instituto Politécnico Nacional (IPN) en 2014.

Actualmente es estudiante del programa de Maestría en Ciencias de la Computación en el Centro de Investigación en Computación del IPN, en el Laboratorio de Ciencia de Datos y Tecnología de Software, interesado en temas como DataMining, Information Visualization, DSS System, GIS System y Big Data.

Realizó una estancia de investigación en la Technische Universität Darmstadt (Universidad Técnica de Darmstadt), en el laboratorio de Visual Analysis and Search Group, bajo la supervisión de la Dr. Tatiana Von Landesberger, Darmstadt, Alemania. Esto con el objetivo en su proyecto de investigación de tesis de Maestría. Esta investigación tiene como propósito el desarrollo de una plataforma web que permita entender los fenómenos urbanos de la Ciudad de México.

Ing. Pedro Ricardo Ortega Castellanos



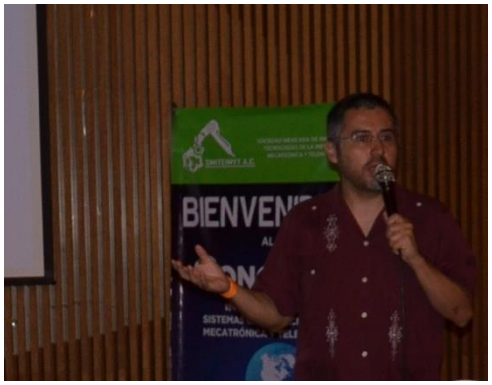
Es ingeniero en Informática por la Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales Y Administrativas (UPIICSA) del IPN y actualmente estudia el último semestre de la maestría en ciencias de la computación del CIC-IPN.

En el Centro de Investigación en Computación se encuentra desarrollando una herramienta informática para el análisis de grandes volúmenes de noticias digitales publicadas en México; implementando técnicas de crawleo en páginas web; recuperación de la información basado en espacios vectoriales; procesamiento de lenguaje natural para reducción de dimensionalidad, extracción de entidades y su reconocimiento sobre texto plano, así como para el agrupamiento de tópicos mediante el uso de algoritmos de aprendizaje automático; y actualmente experimenta con tecnología Big Data para el procesamiento de texto.

Imparte talleres de Minería de Datos, Descubrimiento de Conocimiento sobre Datos Abiertos, Bases de datos SQL Y NOSQL, lenguajes de programación como: Pyhon, Java, JavaScript, CSS y HTML.

Se encuentra certificado como administrador Linux, diseñador de páginas web, desarrollador Android y administrador de bases de datos MySQL. Ha tomado cursos como Big Data con Apache Spark y Scala impartido por MTA SZTAKI de Hungría, desarrollador Java en UPIICSA y Administrador de proyectos (PMP) de Tecnologías de la Información en el CIC IPN.
En el medio profesional ha trabajado como Ingeniero de pruebas de aplicaciones transaccionales Android en CAME, empresa de financiamientos; ingeniero de requerimientos e implementador de soluciones tecnológicas para PYMES.

Ing. Norberto García Lavanderos



El Ing. Norberto García Lavanderos estudio en la Universidad de Ciencias y Administración (UCAD) actualmente realiza su maestría en Ciencias de la Computación en el Centro de Investigación en computación con el tema de Tesis: "Desarrollo de un sistema de gestión de datos y procesos para un clúster de Big Data", ha colaborado con en los congresos CORE 2015, CORE 2016, CORE 2017 al impartir el curso "Procesamiento de datos utilizando herramientas de Big Data" y en el congreso Ingeniería en Sistemas Computacionales, Mecatrónica y Telemática 2016 al impartir el curso "Aplicando Las Herramientas De Big Data". Coordinación y apoyo para el "Workshop Big Data Processing with Apache Flink" 2015, impartido por colaboradores del centro de cómputo SZTAKI, Hungría. Coordinación y apoyo para el "Workshop Big Data Processing with Apache Spark" 2016, impartido por colaboradores del centro de cómputo SZTAKI.

Asistencia al curso de maestría "Tópicos Avanzados en Sistemas de Bases de Datos y Administración de Información a Gran Escala". Ejercicios de transferir información de un sistema gestor de base datos a un ambiente distribuido utilizando sintaxis tipo SQL, como también alojar información no estructurada en forma de tablas (Apache Hive). Implementación de un clúster con Apache Hadoop con nuevas herramientas para procesamiento de información (Apache Hive, Apache PIG, entre otros). Análisis de comportamiento de la carga de información en gran escala y de forma distribuida. Instalación y configuración de un clúster de sistemas de archivos distribuidos con framework para procesamiento y almacenamiento de Big Data.

En el medio profesional ha desempeñado el puesto de: Gerente de Sistemas Informáticos, Ingenieros y Arquitectos ORSI S.A. de C.V. Mantenimiento de computo, Ingeniería Técnica VEGA, S.A. DE C.V. Auxiliar Técnico, Servicios Educativos Integrados al Estado de México.

REQUERIMIENTOS TÉCNICOS

CPU: Intel Core i5 o superior, AMD A6 o superior
RAM: 8 GB
Disco Duro: 128 GB disponibles
Conectividad: FastEthernet o superior, IEEE 82.11n o superior
SO: Linux Ubuntu 16.04

ESTRUCTURA DEL DIPLOMADO

Para lograr el objetivo el Diplomado se encuentra dividido en cuatro módulos y que son:

MODULO A. Introducción a Herramientas Básicas (40 hrs.)

Ejercitar y evaluar al alumno del diplomado en la administración de Linux, en la programación de Java, Python y en la manipulación de bases de datos SQL y NoSQL.

1. Administración-Instalación de Linux (Ubuntu)
   - Comandos del sistema
   - Scripts (shells)

2. Programación
   - Java (Inlellij IDEA)
   - Python
   - Instlación de API-Graphview,
3. Bases de datos SQL
   - MySQL
   - Instalación y configuración
   - MySQL Workbench
   - Carga de bases de datos
   - SQL Consultas (descripción de bases de datos)
   - Conexión con Python, Java
4. Bases de datos NoSQL
   - MongoDB
   - Instalación
   - Robomongo
   - Carga de bases de datos
   - Consultas
   - Conexión Python, Java
5. Evaluación

MODULO B. Proceso de descubrimiento de conocimiento en conjuntos de datos (48 hrs.)

1. ETL-OLAP
   - Conjunto de datos abiertos (Servicios de Salud de la CDMX, ECOBICI, Verificentros, etc)
2. Técnicas básicas de Minería de Datos
   - Patrones frecuentes
   - Generación de Reglas de Asociación
   - Clasificación
   - Clustering
3. Recuperación de Información
   - Modelos de espacio vectorial
   - Modelos de similitud
   - Escalamientos multidimensional
4. Visualización de la información
5. Estructura de la información
   - Análisis de Redes Sociales
6. Evaluación.

MODULO C. Procesamiento de datos con Apache® Spark®. (48 hrs.)

El objetivo del módulo es proporcionar una introducción a la computación distribuida, así como un profundo conocimiento práctico de una de las herramientas más potentes existentes Apache® Spark®. Los estudiantes podrán escribir rápidamente aplicaciones en Java y Python utilizando bibliotecas de Apache® Spark® con conjuntos de datos estructurados, no estructurados y flujos de datos.

1. ¿Qué es Big Data? Desafíos, conceptos, herramientas.

- Introducción al cómputo distribuido. Historia (Hadoop® Framework)
    - Arquitecturas. Planificación de infraestructuras, problemas.
    - Conceptos de gestión de un cluster.
    - YARN (Hadoop 2.+).
- Frameworks, diversos modelos de computación, dominio de problemas.
- Almacenamiento (HDFS).
    - Introducción al marco más popular Hadoop.
    - Información general del paquete Hadoop.
    - ¿Cómo instalar Hadoop? Herramientas necesarias. o Scripts. o Configuración.
- Map-Reduce en Hadoop (YARN).
    - Interfaz web HDFS y YARN.
    - ¿Cómo enviar una solicitud al cluster? (práctico)
    - Protocolo y fases de presentación del YARN. (teórico)
    - ¿Cómo supervisar y recuperar los registros relacionados con una aplicación? (práctico)

2. Panorama del paradigma Map-Reduce (teórico y práctico)

- El paradigma Map-Reduce. El programa Map-Reduce. Fases.
    - Compruebe la API de Map-Reduce (implementación) de Hadoop.
    - Conteo de palabras en un conjunto de datos grande.
    - Concepto de WordCount (cómo deben funcionar los mapas y los reductores).
- Tómese su tiempo con la ejecución, la supervisión, la gestión de aplicaciones (matar, verificar los registros)
- Otros componentes útiles: Partitioners, Combiners, Compression, Counters, Zero Reduces.

3. Introducción a Apache® Spark® (teórica y práctica)

- Arquitectura y comparación con otros sistemas
    - RDD
    - Conceptos básicos de Python
- Ejecutar los siguientes ejemplos:
    - WordCount
    - Detección de Bi-gram.
    - Palindromos
- Instalación, configuración, monitorización o Ejecución y escritura de programas sencillos interactivamente
- Interior de Apache® Spark® (teórico y práctico)
    - Ampliación de los RDD.
    - Linaje de un programa Spark®.
    - Etapas, particiones y tareas.
    - Programación y ejecución.
    - Rendimiento y ajuste fino.
    - Ampliación de la API de RDD.
    - Extensiones sencillas del iterador de Python.
    - Envolver funcionalidades comunes.
    - Creación de un nuevo RDD.

- Envolver una gestión avanzada y común de problemas de manipulación de excepciones en el procesamiento de datos.

4. Trabajos de avanzados con Apache® Spark®: exploración de datos (práctica)

- Introducción al aprendizaje automático
- ¿Qué es el aprendizaje automático? (teoría)
- Vectores (teoría + codificación) K-means
    - Motivación: Detección de anomalías (teoría)
    - K-means (teoría)
    - Cargando los datos de la muestra (práctica)
    - Implementación de K-means (práctica)
    - Uso de MLlib K-means (práctica)
- Spark® SQL (práctico)
- Conceptos SQLContext y DataFrame

5. Evaluación

MODULO D. Procesamiento de datos con Apache® Flink®. (48 hrs.)

El interés en Apache® Flink® superó a Apache® Hadoop®, ya que proporciona una forma más cómoda y rápida de procesar grandes conjuntos de datos y que actualmente se utiliza en una amplia gama de organizaciones para crear grandes aplicaciones de datos. El objetivo del curso es proporcionar una introducción a la computación distribuida, así como un conocimiento práctico en profundidad para el motor analítico general más potente, Apache® Flink®. Los estudiantes podrán escribir aplicaciones rápidamente en Java, Scala mediante el uso de las bibliotecas sofisticadas de Flink® para procesar datos no estructurados, estructurados, grafos de datos o flujos de datos.

1. Introducción a Flink

- ¿Qué es Apache Flink? (teórico)
    - Arquitectura de Flink (comparación con Hadoop y otros sistemas)
- Configuración de Flink
    - Instalación y configuración (práctica)
    - Ejecución de un trabajo simple (por ejemplo, WordCount) (práctico)
    - Interfaz Web, recuperación de registros (prácticos)
    - Configuración de Flink en YARN y ejecución de un trabajo (práctico)
- Programación de Flink
- Operadores de Flink (teóricos)
- Creación de un proyecto Maven con los arquetipos Flink (prácticos)
- Utilizar el IDE para depurar (práctico)
- Pequeños ejercicios con operadores básicos (prácticos)
    - Operadores, Tuplas, estableciendo paralelismo
    - Visualizador del Plan Flink
- Monte Carlo pi estimación (ejercicio práctico)
- Ejecución de la estimación pi en un cluster (práctico)

2. Ejercicios Flink

- Lectura / escritura de archivos HDFS (prácticos)
- Revisando WordCount (práctico)
- Configurar el registro (práctico)
- Lectura / escritura de archivos CSV (prácticos)
- Unir estrategias de operador (teóricas)
    - POJOs, base de datos como operadores (prácticos)
    - Calificaciones de usuario topk (práctica)
- Iteraciones a granel (teóricas)
    - Pequeños ejercicios de iteración (por ejemplo, aumentar números) (prácticos)
    - PageRank (práctico)
    - iteraciones delta (teóricas)
    - Ejercicios pequeños (por ejemplo, propagación mínima del gráfico) (práctico)
    - Componentes conectados (prácticos)
- Kmeans (práctico)

3. Características internas de Flink

- La serialización de Flink (teórica)
- Uso de la serialización Flink (práctico)
- ¿Cómo funciona la capa de tiempo de ejecución de Flink? (teórico)
- Uso de la capa de ejecución para tareas sencillas (prácticas)
- Gestión de la memoria flink (teórica + práctica)
- Programación de trabajos (teórica)
- Comprobación del ciclo de vida del job(práctico)
    - ¿Qué sucede si un job falla?

4. Introducción a ventas de datos (Streaming)

- ¿Qué es streaming? (teórico)
    - Conceptos, desafíos
    - Comparación con otros sistemas de streaming
- Escribir programas de streaming simples (prácticos)
- Conexión a diferentes fuentes (sources) / repositorios (sinks) (por ejemplo Kafka) (práctico)
- Conexión a la API de Twitter (práctico)
    - ¿Cuál es la palabra más frecuente? (práctico)
- Características de ventanas (teórica)
- Ejercicios de ventanas inicial hasta avanzado (práctico)
    - Número de tweets cada 5 minutos
- Disparos de ventanas / política de desalojo (teórico + práctico)

5. Evaluación

BIBLIOGRAFÍA:

Modulo A (Introducción a Herramientas Básicas).

- Thomas, K. (2009). Ubuntu: pocket guide and reference. Place of publication not identified: MacFreda Pub. ISBNs: 9781440478291
- Bautts, T., Dawson, T. & Purdy, G. (2005). Linux network administrator's guide. Sebastopol, Calif: O'Reilly. ISBN: 0596005482
- Urma R. (2015). Introducing Java 8, Raoul-Gabriel Urma. USA:O'Reilly, ISBN: 978-1-491-93434-0
- Introduction to Programming Using Java, David J. Eck Hobart, William Smith Colleges
- Lambert, K. & Osborne, M. (2010). Fundamentals of Python : from first programs through data structures. Boston, Mass: Course Technology/Cengage Learning. ISBN-10: 1-4239-0218-1
- Silberschatz, A., Korth, H. & Sudarshan, S. (2011). Database system concepts. New York: McGraw-Hill. ,ISBN: 978-0-07-352332-3
- Chodorow, K. (2013). MongoDB : the definitive guide. Beijing: O'Reilly. ISBN: 978-1-449-34468-9

Modulo B (Proceso de descubrimiento de conocimiento en conjuntos de datos & Ciencia de Datos)

- Han, J. & Kamber, M. (2006). Data mining : concepts and techniques. Amsterdam Boston San Francisco, CA: Elsevier Morgan Kaufmann.
- Silberschatz, A., Korth, H. & Sudarshan, S. (2011). Database system concepts. New York: McGraw-Hill.
- Zhang, J. (2007). Visualization for information retrieval. Berlin London: Springer.

Modulo C (Procesamiento de Big Data con Apache Spark)

- Turkington, G. (2013). Hadoop Beginner's Guide. Birmingham: Packt Pub.
- White, T. (2012). Hadoop : the definitive guide. Beijing: O'Reilly.
- Holmes, A. (2012). Hadoop in practice. Shelter Island, NY: Manning.
- Perera, S. & Gunarathne, T. (2013). Hadoop MapReduce cookbook : recipes for analyzing large and complex datasets with Hadoop MapReduce. Birmingham: Packt Pub.
- Arora, A. & Mehrotra, S. (2015). Learning YARN : moving beyond MapReduce--learn resource management and big data processing using YARN. Birmingham, UK: Packt Publishing.
- Karau, H., Konwinski, A., Wendell, P. & Zaharia, M. (2015). Learning Spark. Beijing Sebastopol: O'Reilly.
- Sankar, K. & Karau, H. (2015). Fast data processing with Spark : perform real-time analytics using Spark in a fast, distributed and scalable way. Birmingham, UK: Packt Publishing.
- Ramamonjison, R. & Lee, D. (2015). Apache spark graph processing : build, process, and analyze large-scale graphs with Spark. Birmingham, UK: Packt Publishing.
- Richert, W. & Coelho, L. (2013). Building Machine Learning Systems with Python. Birmingham: Packt Publishing.

Modulo D (Procesamiento de Big Data con Apache Flink)

- Introduction to Apache Flink: Stream Processing for Real Time and Beyond, Ellen Friedman and Kostas Tzoumas, O'Reilly, ISBN: 978-1-491-97393-6
- WADKAR, S. (2017). FLINK IN ACTION. S.l: O'REILLY MEDIA.
- Tanmay. (2017). Learning Apache Flink. City: Packt Publishing.