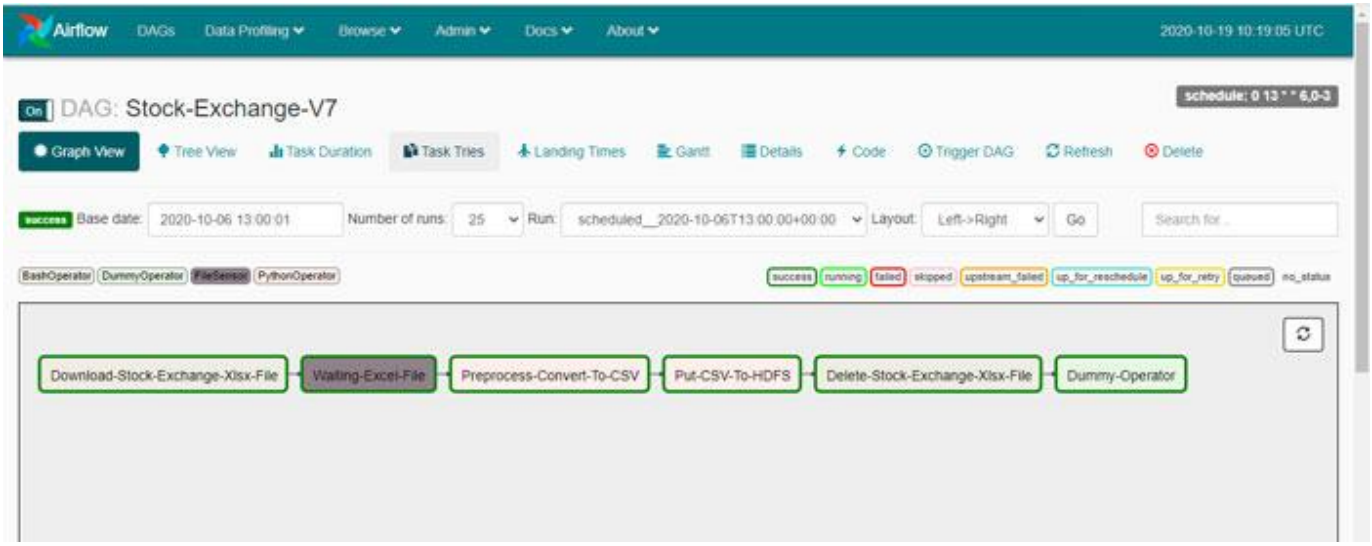
سرفصل‌های این وبکست آموزشی

- آشنایی با تعاریف و مفاهیم پایه مهندسی داده (۳۰ دقیقه)
- انجام مثال عملی پردازش داده‌های روزانه بورس ایران (حدود یک ساعت و ربع)

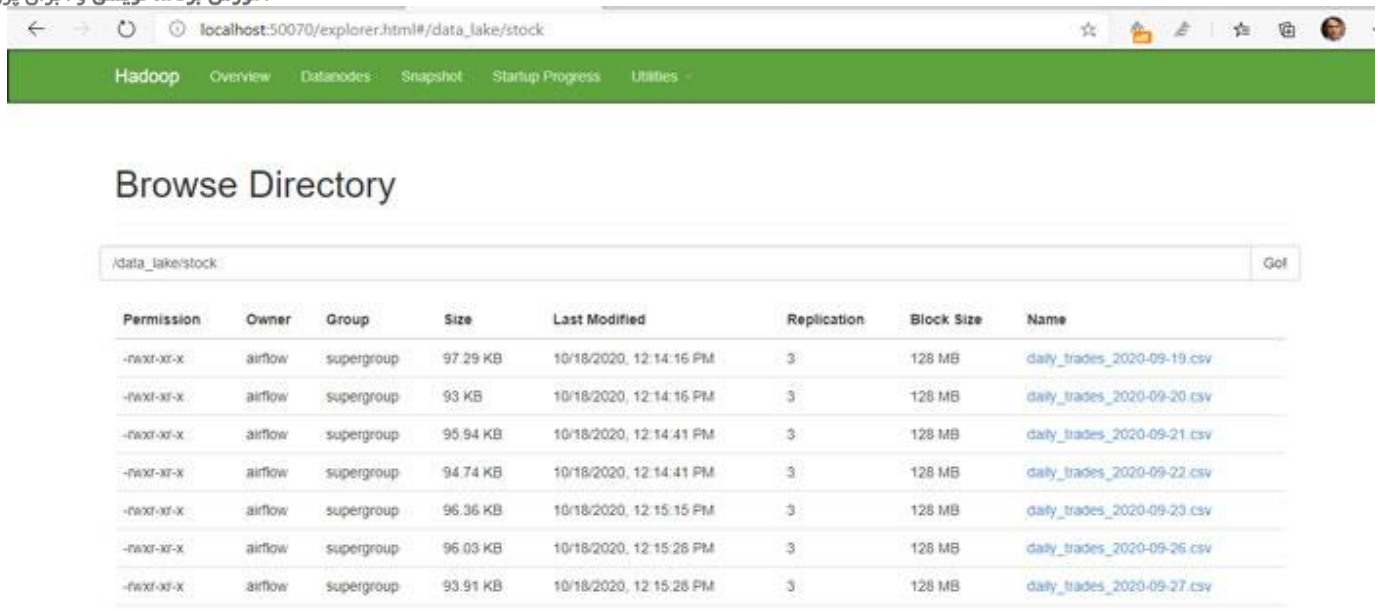
در این مثال کاربردی، داده‌های روزانه بورس ایران از سایت tsetmc به صورت خودکار و روزانه دانلود شده، بعد از پردازش اولیه و تبدیل قالب آن از اکسل به csv ، درون HDFS (سیستم فایل هدوپ) ذخیره می‌گردد. سپس با تعریف جداولی در Hive و با استفاده از Hue ، انواع کوئری های تحلیلی مانند بررسی شرکتهایی که نرخ خرید حقوقی به حقیقی آنها بیشتر از ۲ است، شرکتهایی که حجم معاملات روزانه‌شان از میانگین سه ماهه آنها بیشتر است و ... را روی این داده اجرا خواهیم کرد. بنابراین در مثال اول، با فناوریهای زیر آشنا خواهیم شد:

- Apache Airflow به عنوان محور کار و هماهنگ کننده اصلی تسک‌های روزانه
- Hadoop برای ذخیره و اجرای کوئری های Hive
- Hive به عنوان یک موتورپردازشی SQL بر روی هدوپ و داده‌های حجیم.
- Hue به عنوان یک محیط گرافیکی کار با Hive و اجرای کوئری‌ها

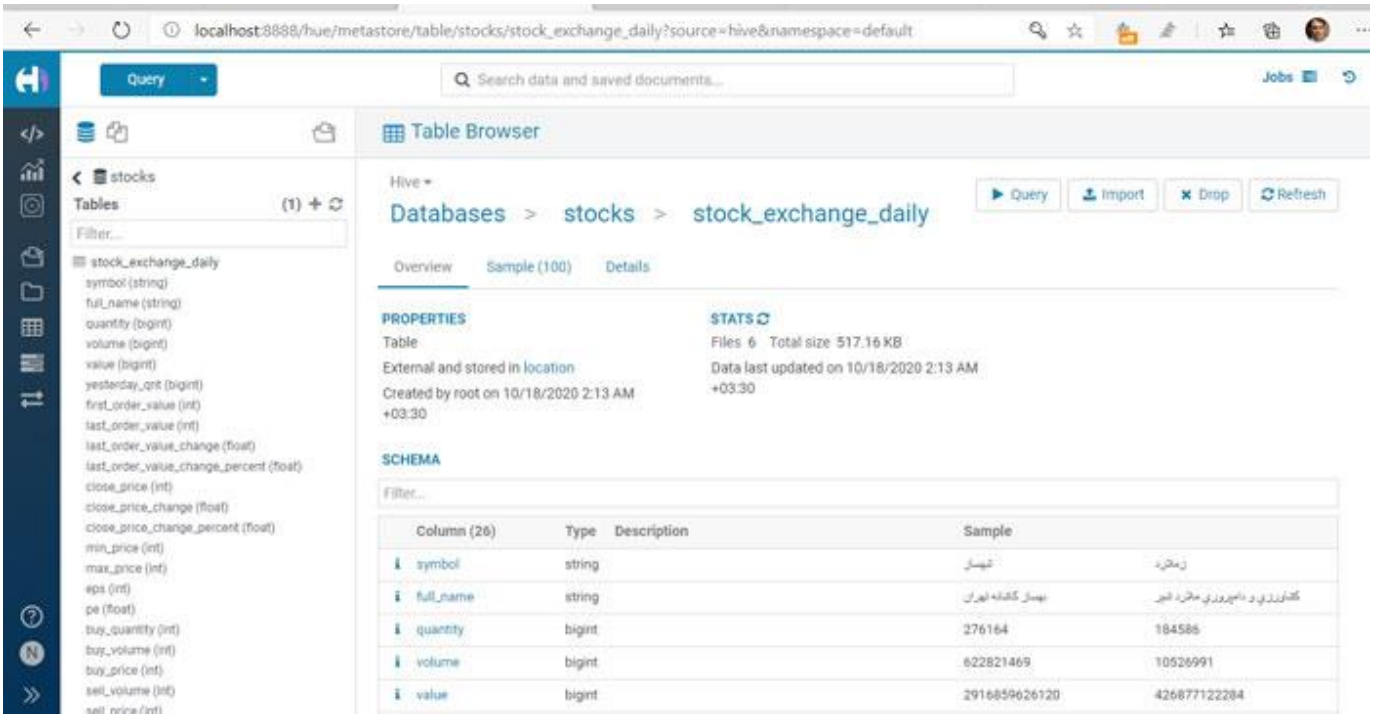
اسکرین‌شات‌هایی از مراحل مختلف انجام این مثال عملی را در زیر می توانید مشاهده کنید:



شکل ۱- نمونه‌ای از خط پردازش داده ایجاد شده در Airflow



شکل ۲- نمونه ای از فایلهای دانلود شده روزانه بورس درHDFS



شکل ۳ – تصویری از محیط Hue که جدول ایجاد شده در Hive را نشان میدهد

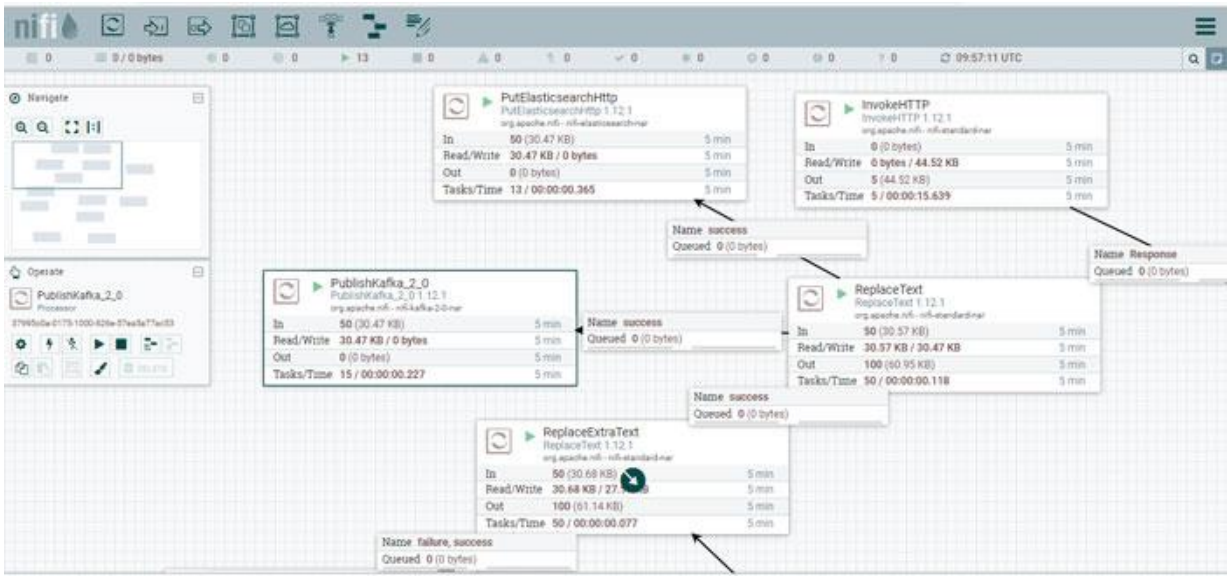
- انجام مثال عملی پردازش توئیت‌های لحظه‌ای سایت سهامیاب (حدود یک ساعت و ربع)

سایت سهام‌یاب به عنوان یکی از مراجع اصلی کاربران بورس ایران، روزانه توئیت‌ها و مطالب مختلفی را از کاربران دریاف می‌کند. در مثال دوم این کارگاه عملی، سعی می‌کنیم این توئیت‌ها را دریافت کرده، پردازش نماییم و نهایتا آنها را ذخیره کنیم و علاوه بر ساخت داشبوردهای تحلیلی با الاستیک سرچ، به کمک اسپارک و امکانات پردازش جریان آن، اطلاعات مفیدی راجع به هر نماد بورسی در لحظه استخراج نماییم. در این مثال با ابزار زیر کار خواهیم کرد

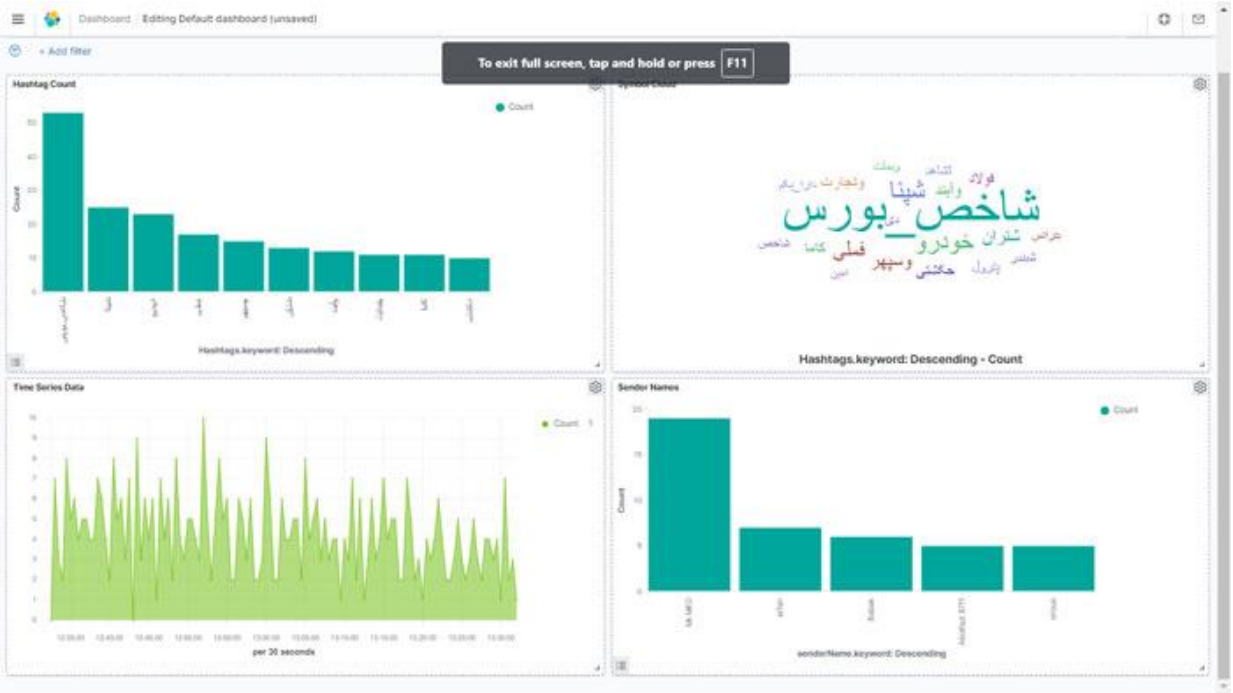
:

- Apache Nifi: به عنوان محور اصلی کار که طراحی خط پردازش داده، ما بر اساس آن انجام خواهد شد.
- Kafka به عنوان صف توزیع توئیت‌ها در شبکه
- Elastic Search: برای ذخیره توئیت‌ها
- Kibana: برای ساخت داشبوردهای لحظه‌ای
- Spark: برای پردازش توئیت‌ها و استخراج آمار مورد نیاز که از نسخه جدید اسپارک (نسخه ۳) و رهیافت نوین Spark Structured Streaming استفاده خواهد شد.

در زیر اسکرین‌شاتهایی از ابزار مورد استفاده و خروجی ها می توانید مشاهده کنید:



شکل ۴- تصویری از محیط آپاچی نایفای برای پردازش لحظه‌ای توئیت‌های بورسی



شکل ۵- داشبورد اولیه طراحی شده برای پایش توئیت‌های لحظه‌ای بورس به کمک کیبانا

شکل ۶ – مشاهده توئیت‌های دریافتی به صورت لحه‌ای در آپاچی کافکا

در مثال‌های عملی این وبنیار و دوره آموزش مقدماتی مهندسی داده، از داکر برای ساخت کلاسترهای پردازشی استفاده خواهیم کرد که با رم ۴ گیگابایت، تست شده و آماده استفاده هستند.

مجتبی بنائی [مشاوره، مدرس و متخصص مهندسی داده]

دانشجوی دکترای نرم‌افزار دانشگاه تهران، مدرس دانشگاه و فعال در حوزه مهندسی نرم‌افزار و علم داده که تمرکز کاری خود را در چند سال اخیر بر روی مطالعه و تحقیق در حوزه کلان‌داده و تولید محتوای تخصصی و کاربردی به زبان فارسی و انتشار آنها در سایت مهندسی داده گذاشته است. مدیریت پروژه‌های نرم‌افزاری و طراحی سامانه‌های مقیاس‌پذیر اطلاعاتی از دیگر فعالیتهای صورت گرفته ایشان در چند سال گذشته است.

- مدرس دوره‌هایBigData ی نیک آموز
- دانشجوی دکترای نرم افزار و مدرس دانشگاه تهران
- مجری و مشاور پروژه‌های کلان‌داده در سطح ملی و بین المللی
- فعال در حوزه تولید محتوای تخصصی در زمینه پردازش داده